DOCUMENT RESUME

ED 198 157 TM 810 100

AUTHOR	Bulcock, J. W.: And Others
TITLE	The Collinearity Free and Bias Reduced Regression Estimation Project: The Theory of Normalization Ridge Regression. Report No. 2.
INSTITUTION	Alberta Univ., Edmonton. Dept. of Educational Foundations.
SPONS AGENCY	Natural Sciences and Engineering Research Council, Ottawa (Ontario).
PUB DATE	Nov 80
GRANT	NSERC-A3960
NOTE	163p.
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	*Comparative Analysis: *Mathematical Models: Measurement Techniques: *Theories
IDENTIFIERS	*Multicollinearity: *Normalization Regression Estimation: Ridge Regression Analysis

ABSTRACT

Multicollinearity refers to the presence of highly intercorrelated independent variables in structural equation models, that is, models estimated by using techniques such as least squares regression and maximum likelihood. There is a problem of multicollinearity in both the natural and social sciences where theory formulation and estimation is in the mathematical model building tradition. The variety of ways in which analysts can cope with the condition of multicollinearity is presented. The measure of multicollinearity developed is a nonlinear transformation of the maximum variance inflation factor (Vmax) which, like the correlation coefficient, gives zero for no collinearity and one for perfect collinearity. Three ridge regression estimating methods are described and their advantages and disadvantages compared. The first three disadvantages of ridge regression stem from the use of the total mean square error (MSE) criterion. The last two chapters of the report address the theory and practice of normalization ridge regression analysis. It is shown that if a new criterion called the variance normalization criterion is used, several of the inherent disadvantages based on using the MSE criterion are correctable. Results indicate that normalization ridge regression analysis has advantages (in terms of the performance indices) over seven other methods. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED198157

REPORT NO. 2
GRANT NO. A3960

THE COLLINEARITY FREE AND BIAS REDUCED REGRESSION ESTIMATION PROJECT

THE NATURAL SCIENCES AND ENGINEERING
RESEARCH COUNCIL OF CANADA

THE THEORY OF NORMALIZATION RIDGE REGRESSION

by

J.W. BULCOCK
W.F. LEE
W.S. LUK

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. W. Bulcock

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM810100

NOVEMBER 1980

NOTICE

This report is submitted for review according to the procedures followed by the Natural Sciences and Engineering Research Council of Canada.

The work on which this report is based was performed under NSERC Grant No. A3960 awarded the first author, April 1980.

Copies of the report are available from:

Department of Educational Foundations
University of Alberta
Edmonton, Alberta
T6G 2G5

THE AUTHORS

Jeffrey W. Bulcock*
Honorary Visiting Professor
Department of Educational Foundations
University of Alberta
Edmonton, Alberta
T6G 2G5

* On leave from:

Department of Educational Foundations
Memorial University of Newfoundland
St. John's, Newfoundland
A1B 3X8

Wan-Fung Lee
Department of Community Medicine
Faculty of Medicine
Memorial University of Newfoundland
St. John's, Newfoundland
A1C 5S7

Wo-Shun Luk
Computing Science, Rm 7316
Simon Fraser University
Burnaby, B.C.
V5A 1S6

CONTENTS

CHAPTER PAGE

0.1 Report Summary 1

CHAPTER 1. THE PROBLEM OF MULTICOLLINEARITY........ 1

1.1 Multicollinearity.................. 8

1.2 The Effects of Multicollinearity on OLS Regression Procedures 14

1.3 Measures of Multicollinearity.... 21

1.4 Standard Solutions................ 25

1.4.1 The random error.................. 25

1.4.2 The sample size, n................ 26

1.4.3 The degree of multicollinearity.............................. 26

1.5 Summary and Overview.............. 27

CHAPTER 2. SIMPLE RIDGE REGRESSION IN THEORY........ 29

2.1 Simple Ridge Regression........... 29

2.2 The Properties of Simple Ridge Regression 32

2.3 The Optimal k 50

CHAPTER 3. SIMPLE RIDGE REGRESSION IN PRACTICE 52

3.1 Hoerl and Kennard's Ridge Trace Method........................ 52

3.2 Kasarda and Shih's Method......... 55

3.3 Vinod's Indix of Stability Method............................... 58

3.4 Advantages and Disadvantages of Simple Ridge Regression........ 60

CHAPTER 4. SIMPLE NORMALIZATION RIDGE REGRESSION ANALYSIS.................................. 65

4.1 Introduction........................ 65

4.2 The Normalization Criterion....... 67

CHAPTER PAGE

4.3 The Normalization Criterion versus the Mean Square Error Criterion 68
4.4 Performance Indices.................. 71
4.4.1 The biasing parameter, k............ 71
4.4.2 Inflation in the residual sum of squares, ϕ.......................... 72
4.4.3 The maximum variance inflation factor, Vmax.......................... 72
4.4.4 The acceptance level of ridge estimate, α 73
4.4.5 The degree of multicollinearity, Dmax................................ 73
4.4.6 The empirical sensitivity ratio, E.S.Ratio 73
4.4.7 Relative reduction in variance, RVAR.................................... 74
4.4.8 The average percentage error of estimated variance, AV.%Er. VAR.................................... 74
4.4.9 The coefficient of determination, R^2............................... 75

CHAPTER 5. RIDGE REGRESSION AND NORMALIZATION RIDGE REGRESSION: SOME COMPARISONS........ 76

5.1 Introduction.......................... 76
5.2 The Models 78
5.2.1 The ten-factor example from Hoerl and Kennard (1970b)........... 78
5.2.2 The fifteen-factor air pollution problem from McDonald and Schwing (1973)......... 79
5.2.3 The ten-factor gas mileage study from Hocking (1976)........... 81
5.2.4 The five-factor employee satisfaction example from Price (1977)................................. 81
5.3 The Estimates......................... 86
5.4 Performance Comparisons............. 95

CHAPTER PAGE

5.5 Ridge Trace Results 111
5.6 Conclusions 111

CHAPTER 6. CONCLUSION 121

REFERENCES 125

APPENDIX A. Program SNR 130
A.1 Description 130
A.2 Program Listing 131

LIST OF TABLES

Table	Description	Page
5.1	Correlation Matrix and Eigenvalues of X'X and X'y for the Ten-Factor Example from Hoerl & Kennard (1970b) (N=36)	80
5.2	Description of Variables in the Fifteen-Factor Air Pollution Problem from McDonald & Schwing (1973)	82
5.3	Correlation Matrix and Eigenvalues of X'X and X'y for the Fifteen-Factor Air Pollution Problem from McDonald & Schwing (1973) (N=60)	83
5.4	Description of Variables in the Ten-Factor Gas Mileage Problem from Hocking (1976)	84
5.5	Correlations and Eigenvalues of X'X and X'y for Hocking's (1976) Gas Mileage Problem (N=32)	85
5.6	Correlation Matrix (X'X) and Eigenvalues for X'y for the Five-Factor Employee Satisfaction Example from Price (1977) (N=30)	87
5.7	Standardized Regression Coefficients and their t-values (in parentheses) for the 10-Factor Example from Hoerl & Kennard (1976b)	90
5.8	The Significant Parameters (✓) in the Hoerl & Kennard 10-Factor Example by Estimating Method	91
5.9	Standardized Regression Coefficients and their t-values (in parentheses) for the 15-Factor Air Pollution Problem from McDonald & Schwing (1973)	93
5.10	The Significant Parameters (✓) in the McDonald and Schwing 15-Factor Air Pollution Problem by Estimating Method	94
5.11	Standardized Regression Coefficients and their t-values (in parentheses) for the 10-Factor Gas Mileage Problem from Hocking (1976)	96

LIST OF TABLES

Table	Description	Page
5.12	The Significant Parameters (✓) in the Hocking 10-Factor Gas Mileage Problem by Estimating Method	97
5.13	Standardized Regression Coefficients and their t-values (in parentheses) for Price's (1977) Employee Satisfaction Problem	98
5.14	The Significant Parameters (✓) in the Price 5-Factor Employee Satisfaction Study by Estimating Method	99
5.15	The Characteristics and Performance Indices of OLS and Ridge Regression Procedures for the 10-Factor Example from Hoerl & Kennard (1970b) (p=10, N=36, Dc=.952)	103
5.16	The Characteristics and Performance Indices of OLS and Ridge Regression Procedures for the 15-Factor Air Pollution Problem from McDonald & Schwing (1973) (p=15, N=60, Dc=.9970)	104
5.17	The Characteristics and Performance Indices of OLS and Ridge Regression Procedures for the 10-Factor Gas Mileage Problem from Hocking (1976) (p=10, N=32, Dc=.985)	105
5.18	The Characteristics and Performance Indices of OLS and Ridge Regression Procedures for the 5-Factor Employee Satisfaction Study (Price, 1977) by Estimating Method (p=5, N=30, Dc=.985)	106
5.19	Rankings of Performance Indices for Five Ridge Regression Methods	108

LIST OF FIGURES

Figure	Description	Page
2.1	The Geometrical Relationship of the OLS Estimate, the Ridge Regression Estimate, and the Inflation in the Residual Sum of Squares....................	33
3.1	The Ridge Trace of the 10-Factor Problem from Hoerl and Kennard (1970b)..	53
3.2	The Mean Square Error Function (MSE) and its OLS estimate (MSE) where k_s is Kasarda & Shih's k, and k_o is the true optimum k...............................	57
3.3	The Multiple Minimum of ISRM for the 6-Factor Example from Vinod (1976)........	61
5.1	Ridge Trace Result for the 10-Factor Example from Hoerl & Kennard (1970b) (N=36, Vmax=9.11, Dmax=.952)..............	112
5.2	Ridge Trace Result for the Air Pollution Problem from McDonald & Schwing (1973) (N=60, Vmax=104.2, Dmax=.994).....................................	113
5.3	Ridge Trace of the 10-Factor Gas Mileage Problem from Hocking (1976) (N=32, Vmax=21.79, Dc=.985)................	114
5.4	Ridge Trace Result for the 5-Factor Employee Satisfaction Problem from Price (1977) (N=30, Vmax=421.8, Dmax=.998).....................................	115

0.1 REPORT SUMMARY

The Problem of Multicollinearity

Multicollinearity refers to the presence of highly intercorrelated independent variables in structural equation models; that is, models estimated by using techniques such as ordinary least squares regression, generalized least squares, two-stage least squares, three-stage least squares, and full-information maximum likelihood. Its effect is to invalidate some of the basic assumptions on which these estimating methods are based; and, hence, to render inappropriate the use of these procedures, unless modified. The problem of multicollinearity is ubiquitous -- that is, invariant in degree -- in both the natural and social sciences where theory formulation and estimation is in the mathematical model building tradition. Thus, multicollinearity has been extensively studied in technometrics, biometrics, econometrics and psychometrics. The problem is particularly acute in the social sciences where variables seldom present themselves in the pure or unidimensional form as for example time or mass in physics. The variety of ways in which analysts in the structural equation tradition can cope with the condition of multicollinearity is the subject of this report.

The Effects of Multicollinearity

The six harmful effects of multicollinearity are as follows: (i) the estimated regression coefficients will have large sampling variance; (ii) the estimated coefficient vectors will be too long; (iii) models estimated using traditional techniques will not give accurate prediction even when the goodness of fit or the coefficient of determination seems high; (iv) the beta estimates will be sensitive to sampling fluctuation; (v) the estimated regression coefficients will be highly dependent, hence error prone; and (vi) the regression estimates will be highly sensitive to model mispecification.

Measures of Multicollinearity

Though there is some consensus that the maximum variance inflation factor (Vmax) is the best single measure of multicollinearity the one developed in this report is a nonlinear transformation of Vmax which, like the correlation coefficient, gives zero for no collinearity and one for perfect collinearity.

Simple Ridge Regression

The several simple ridge regression techniques use the total mean square error (MSE) function as the criterion for measuring the "goodness" of an estimator. From a geometrical point of view MSE represents the expected squared distance between the estimate of an estimator, $\hat{\beta}$, and the true vector, β. Since the best estimator is the one that minimizes this distance the MSE criterion seems to be a viable alternative to that of ordinary least squares (OLS) -- the minimum residual sum of squares criterion. Unfortunately, however, out of the sixteen properties of simple ridge regression three are problematic. The most serious problem is that an optimum value for the ridge parameter, k, can only be calculated if the true regression coefficient vector and the variance of the residual in the linear model are known. But these parameters are population parameters, not universal constants. Because of this dependency the optimum value of the ridge parameter k cannot be calculated. Instead, it has to be estimated from the sample data. Over fifteen estimating methods have been described each with its distinct advantages and disadvantages. None, however, can guarantee to give a better k than any other, and none can guarantee a smaller MSE to that of OLS regression.

Simple Ridge Regression in Practice

Three ridge regression estimating methods are described and their advantages and disadvantages compared. One of the earliest versions of simple ridge regression developed by the founders of the technique, Hoerl and Kennard, was called the ridge trace method. This procedure enabled the analyst to determine the optimum k visually. A second method developed by Kasarda and Shih is based on the argument that since the OLS estimates of the variance of the residual and the beta coefficient has the minimum variance among all unbiased estimates, then the two OLS estimates of these parameters may be used to replace their true values. Then, the optional k may be obtained by minimizing the OLS estimate of MSE. A third method of simple ridge regression has been developed by Vinod. It is based on the notion that since Hoerl and Kennard's concept of stability will be zero for a completely orthogonal system, then the index of stability of relative magnitude (ISRM) should be minimized to reduce the effects of the nonorthogonality of ill-conditioned data.

These methods have the following advantages: (i) compared to OLS regression they produce smaller sampling variance; (ii) because they produce smaller sampling variance ridge estimates are less sensitive to sampling fluctuation; (iii) ridge solutions reduce the variance of the regression coefficients which renders these solutions less sensitive to model misspecification than other methods; (iv) since the higher the variance of the estimates the higher the dependence of the regression coefficients, and the higher the dependence the greater their error, a ridge regression procedure which reduces variance will generate more accurate estimates; (v) the ridge procedure generates a shorter coefficient vector than OLS; and (vi) although ridge procedure always produces a more accurate prediction if the biasing parameter is not too large.

But these compelling reasons for using ridge regression are rendered largely academic when ridge regression is evaluated in terms of its disadvantages. Three of the properties of ridge regression are problematic, and these account for four disadvantages. Since the total MSE of ridge regression depends on two true, unknown population parameters, the total MSE cannot be obtained. For any regression problem there exists a positive "optional k" (k_0) which is (absurdly) true even for an orthogonal system for which the multicollinearity is zero. Because the optimum k depends on two population parameters (the true regression coefficient vector and the variance of the residual of the linear model) and since both in turn are dependent on the y-variable, then k depends on the dependencies between X and Y variables when, in fact, multicollinearity arises solely from the interdependencies among the explanatory variables, X's. Thus, the ridge regression estimate of k is stochastic when in fact (and theory) it should be nonstochastic.

The four dilemmas of ridge regression stem from these three problematic properties.

(i) The theoretical value of k is stochastic when it should be nonstochastic.

(ii) The optional k depends on unknown population parameters, a fact which renders the problem unsolvable.

(iii) Even for an orthogonal system (the boundary condition) there is an optimum k at which MSE is minimum.

(iv) The bias of the ridge estimates cannot be calculated; thus, the performance of ridge regression cannot be evaluated.

Simple Normalization Ridge Regression Analysis

The matter at issue is whether a ridge estimating procedure can be developed which retains the advantages of the standard ridge regression procedure while overcoming its inherent disadvantages. The first three disadvantages of ridge regression stem from the use of the MSE criterion. The last two chapters of this report address the theory and practice of normalization ridge regression analysis. First, it is shown that if a new criterion called the variance normalization criterion is used several of the inherent disadvantages of ridge regression based on the MSE criterion are correctable. Second, application of the new method through estimation of four problem data sets selected from the ridge regression literature shows that the new criterion performs more consistently than the more traditional MSE criterion.

Mathematically it may be shown that in order to perform the task of reducing variance inflation due to multicollinearity, and only multicollinearity, the variance inflation factor (VIF) -- that is, the diagonal elements of the inverse of the simple correlation matrix -- should be normalized such that it is equal to one. Note that the VIF of each term is a measure of the collective impact of the interdependency of the explanatory variables on the variance of the regression coefficient of that term. Thus, its value provides an indication of the number of times the variance has been inflated; where one stands for no collinearity and infinity for perfect collinearity.

The results of a series of comparative analyses in the last chapter of the report are evaluated in terms of the following performance indices: (i) the biasing parameter, k; (ii) the maximum variance inflation factor, Vmax; (iii) the degree of multicollinearity, Dmax; (iv) the empirical sensitivity ratio, E.S.Ratio; (v) the relative reduction in variance, RVAR; (vi) the inflation in the residual sum of squares, ϕ; (vii) the acceptance level of the ridge estimate, α; (viii) the average percentage error of estimated variance, Av.%Er.VAR; and (ix) the coefficient of determination, R^2.

Ridge Regression and Normalization Ridge Regression Comparisons

The results were not interpreted as being conclusive. They indicated clearly, however, that the normalization method and Hoerl and Kennard's iterative method of ridge regression analysis had significant advantages (in terms of the performance indices) over six other methods -- one being OLS regression. The normalization method proved more consistent than Hoerl and Kennard's iterative method. Monte Carlo simulation experiments are called for to facilitate identification of the governing conditions under which each method will generate its optimum performance.

CHAPTER ONE

1.0 THE PROBLEM OF MULTICOLLINEARITY

1.1 Multicollinearity

When the explanatory variables in a model are highly correlated between themselves it becomes difficult for the analyst to disentangle their separate effects on the dependent variable. What happens in the presence of this condition is that parameter estimates become very sensitive to errors, and their estimation becomes impracticable because the variance inflation of the estimates often makes the partial regression coefficients uninterpretable. Such a condition is known as multicollinearity. The variety of ways that analysts can cope with this condition is the subject of this monograph.

The problem which was first identified by the Norwegian econometrician Ragnar Frisch (1934) is not one of kind; that is, of existence or nonexistence, but rather one of degree; that is, the analyst's concern is always with the magnitude or seriousness of multicollinearity. In other words, the problem is ubiquitous in structural model estimation. It refers to a situation in which predictor vectors, or independent variables "X's" depart from orthogonality - a state of affairs in which a degree of linearity exists among the explanatory variables. Perfect multicollinearity, or complete linear dependency, and

18

its obverse, a complete absence of multicollinearity, or linear independence, are rare limiting conditions.

If the interrelationships between the independent variables are of modest proportions it is seldom difficult to disentangle their effects on the dependent variable. In such instances linear dependence will be nonproblematic. There will be no redundant independent variables in the sense that no single vector in a set of vectors will be expressible as a linear combination of the other vectors. Some, however, could well prove superfluous in that their presence in an equation may not significantly increase the equation's explanatory power. Indeed, where the independent variables are relatively orthogonal the purpose of the estimating exercise might well be the identification of superfluous variables; such is particularly the case when the analyst is interested in comparing rival theories.

The trouble is that in the social sciences variables seldom present themselves in a pure or unidimensional form as for example time or mass in physics. More often than not, independent variables in social science models are co-related operationally even though they may be conceptually distinct. Thus, the Dean of American sociological methodology has identified multicollinearity as one of several tough research problems facing the analyst at the interface of theory and research; that is, problems which lie between the conceptualization of a model (in cause and effect statements) and its measurement (in statistical or probabilistic statements) (cf.

Blacock, 1979, p. 881). Blacock (1979) also identifies (i) the specification and estimation of nonrecursive models, and (ii) the linking of micro-and macro-theories involving two or more units of analysis, as additional issues illustrative of the obstacles to the integration of theory and research. But multicollinearity is a common cause of both problems.

The incorporation of feedback linkages in causal models aggravates the multicollinearity problem. This has been illustrated by the authors (Bulcock et.al. 1979, 1980, 1981) elsewhere through the estimation of feedback models of school learning. Unless the analyst is able to cope with collinearity, feedback models are all too frequently uninterpretable. Similarly, the use of the average value of independent variables as measures of contextual effects in mixed, micro-and macro-models; or the use of the average level of a dependent variable as an indicator of the typical behaviour of an individual's membership group, as a contextual independent variable, will often generate high correlations between individual scores and their respective group means. This is especially the case when $\bar{X}$- and $\bar{Y}$-type models are specified for relatively homogeneous groups, because in the presence of this condition the differences between X_{ij} and $\bar{X}_j$ (where X_{ij} represents the behaviour of the ith individual in the jth group and $\bar{X}_j$ is the typical behaviour of the group) will be small compared to the overall variation in X_{ij} across all individuals.

Thus, both of these important research issues - the problems attendant upon the estimation of nonrecursive models, and the problems attendant upon the estimation of micro-macro-models involving two or more unit levels of analysis - are not merely problematic in and of themselves, but also because of the confounding effect of multicollinearity. In fact it is probably not unreasonable to assume that the estimation of models which at one and the same time call for both feedback linkages and micro-macro-level constructs will prove extremely difficult unless the analyst is able to cope first with the multicollinearity problem.

As contemporary societies enter a post-industrial era they become more complex. Increasingly, simple one-to-one, cause-effect, solutions to complex problems prove unsatisfactory; in fact simplistic "solutions" may prove counter-productive. This is the message from the Carnegie Council on Children with regard to public policies in the child welfare realm (Keniston, 1977; de Lone, 1978; Ogbu, 1978) to cite but one example. The increasing complexity of social systems is paralleled by research developments in the social sciences. The path of scientific development has proceeded from the formulation and estimation of simple models to more complex models. The tentative theory of experimental design developed by J. Stuart Mill was formalized two generations later by Sir Ronald Fisher (1925, 1935); and the controlled experimental approach in the social sciences has given way in the past decade to structural equation model building.

Thus, the earliest statistical models were designed to analyze the data of controlled experiments. These involved very few variables since the potential disturbances of contaminating influences were eliminated through randomization. But few social policy questions can be structured to fit the experimental mold. The testing of several treatments necessarily restricts the population to that included in the experimental design, and the experimental subjects are seldom representative of the population of substantive research interest. Thus, the preferred realism of the natural setting is usually absent in controlled experiments.

Kish (1975) points out that "a research design always involves compromises of the desirable with the possible." And because the advantages of experimentation - control through randomization - is at the expense of the representativeness of samples which permits low level inferences to the defined population, compromise has taken the form of alternative methodologies. In the 1960's widespread use was made of ordinary least squares (OLS) estimating techniques as a single equation method for estimating recursive models of the interdependencies among sets of variables. In the 1970's advances centered on the development of nonrecursive single equation methods of which the two-stage least squares (2SLS) method has proven to be the most widely used of a variety of possible limited-information estimators (Intriligator, 1978, pp. 419-420).

Though it is claimed that these advances are being used because most of their known statistical problems have been resolved (Land and Felson, 1978), they have given rise to a new generation of problems (sui generis). For example, even though OLS regression retains the Gauss-Markov property of minimum variance, this does not mean that the variance is the lowest possible variance. The minimum variance can still be inflated due to multicollinearity; hence, unacceptably large. Further, in misspecified models - and few social science models are fully specified in the sense that all the model predictors are known - OLS estimates are biased. This means that sometimes the "small" variance property will not offset the bias because of all estimating techniques OLS has the largest inflated variance when multicollinearity is high (Johnston, 1972, p. 411).

The important point is that the OLS procedure is only theoretically unbiased. In social science practice, where current models are more often than not misspecified, dependencies frequently exist between the endogenous variables (Y_i's) and the residuals (E_i's). Such a condition will lead OLS estimates ($\hat{\beta}_i$'s) to be biased. The bias tends to be higher when there are too few predictors than when there are too many.

On the other hand, 2SLS performs better than OLS both in terms of bias and MSE. Further, it is less sensitive to specification error (that is, to the addition or deletion of an explanatory variable) and measurement error, more consistent, and has the added advantage of being able to estimate non-recursive model parameters. Nevertheless, the 2SLS estimator is very sensitive to multicollinearity (King, 1974; Johnston, 1972); thus, when the Z_i (right hand side endogenous variables) in 2SLS equations are highly intercorrelated some of the resultant parameter estimates may be uninterpretable.

Because multicollinearity is a major stumbling block to the accurate estimation of structural equation models it is obligatory to examine the specific effects of multicollinearity on standard estimating procedures. We begin with an examination of the effects of multicollinearity on ordinary least squares regression methods.

1.2 The Effects of Multicollinearity on OLS Regression Procedures

It is well known that in all the social sciences structural equation model building, which is also known as causal modelling or path analysis, has been extensively used in the past decade. The basic technique in this type of model building is ordinary least squares (OLS) regression. In fact OLS regression is so extensively used in all the sciences, natural as well as social, that it can be described as a standard statistical technique (e.g., McCabe, 1978).

It has been pointed out that one of the most important difficulties with using OLS regression is its sensitivity to multicollinearity. It has also been pointed out that in the social sciences in particular, because of the lack of researcher control over the variables, the use of sets of nearly independent predictor variables is virtually unknown; therefore, multicollinearity is an unavoidable problem. Thus, although the least squares technique is theoretically sound, empirically it is problematic because when a data set has a high degree of multicollinearity the procedure will not produce satisfactory results (Stein, 1955; Hoerl & Kennard, 1970a; Marquardt, 1970). Further, multicollinearity is an intrinsic property of multi-stage, multi-equation, structural models because the better the specification at stage one of a model the lower the probability of linear independence at subsequent stages of the model.

The ill-effect of the problem in the general case will produce estimates with large variances and this in turn leads to estimates that are unreliable (even of the wrong sign) and sensitive to sampling error and model misspecification (Hoerl & Kennard, 1970a). It is, therefore, of the utmost importance for the statistical analyst to realize the existence and seriousness of the problem before making any claim based on statistical inferences.

The classical multiple regression model is given by:

$$y = X\beta + \varepsilon \qquad (1.1)$$

where y is an (n x 1) vector of observations on the dependent variable, X is an (n x p) matrix of observations on p explanatory variables; where for convenience X and y are scaled so that X'X and X'y give the correlation matrices; and where β is the (k x 1) vector of regression coefficients and ε is an (n x 1) vector of the random error of y.

In OLS estimation it is assumed:

(i) that the X's are nonstochastic, though as Johnston (1972, pp. 267 ff.) notes this assumption can be relaxed to accommodate stochastic variables;

(ii) that there are no (exact) linear relationships between the explanatory variables; that is, X is of full rank; and

(iii) that the error terms have independent normal distribution with zero mean and constant variance; that is, $\varepsilon \sim \text{NID}(0,\sigma^2)$.

The robustness, and the effects of the violation of these assumptions have been thoroughly discussed by many researchers in numerous advanced texts. Here, concern focuses on the second assumption. Mathematically, insofar as there is no exact linear relationship between any two predictors, the model can be solved. But from a practical point of view this is not enough. If a nearly linear relationship exists among the predictors the OLS "solution" will usually prove unacceptable because such a "solution" will be unstable, unreliable, and difficult, if not impossible, to interpret. These are symptoms of the multicollinearity problem. In practice in the social sciences most predictor variable sets are intercorrelated; thus, the problem is ubiquitous, and always a matter of degree.

The OLS solution to the regression problem may be written as:

$$\hat{\beta} = (X'X)^{-1}X'y \tag{1.2}$$

with the variance-covariance matrix,

$$\text{cov}(\hat{\beta}) = \sigma^2(X'X)^{-1} \tag{1.3}$$

When there is a high degree of multicollinearity, the determinant $|X'X|$ will be very small and $(X'X)^{-1}$ will have very large entries (in the extreme case, $|X'X| = 0$ and $(X'X)^{-1}$ would approximate infinity), therefore, each estimated coefficient would have a very large variance and covariance, which is the case for all the problems associated with multicollinearity as discussed below.

When a data set has a high degree of multicollinearity it is characterized by the fact that the smallest eigenvalue of its correlation matrix (X'X) is much smaller than unity (Hoerl & Kennard, 1970a). This condition generates at least six harmful effects of the OLS estimates of β.

1.2.1 *The estimated regression coefficients will have very large sampling variance.*

The total sampling variance for OLS estimates is given as follows (Hoerl & Kennard, 1970a):

$$\Sigma \mathrm{Var}(\hat{\beta}) = \sigma^2 \mathrm{Tr}(X'X)^{-1} \tag{1.4}$$

$$= \sigma^2 \Sigma \frac{1}{\lambda_i} > \sigma^2/\lambda_{min} \tag{1.5}$$

where, and hereafter, Σ represents the summation from i = 1 to i = p, the number of predictors, and λ_{min} is the smallest eigenvalue of the correlation matrix (X'X), which approaches zero for data sets approaching perfect collinearity. Therefore, the total variance; and, hence, the sampling variance of the individual predictors can be very large and may approach infinity when the data set approximates perfect collinearity.

From the geometrical point of view, the total variance can be regarded as the squared distance from the estimated coefficient vector, $\hat{\beta}$, to the true vector β, and the large total variance means that the distance from $\hat{\beta}$ to β is too long; hence, the greater the collinearity the longer the $\hat{\beta}$ to β distance.

1.2.2 *The estimated coefficient vector $\hat{\beta}$ is far too long in general for data sets with a high degree of multicollinearity.*

In general the expected square length of the estimated coefficient vectors $\hat{\beta}$ are far too long in ill-conditioned data sets as Hoerl & Kennard (1970a) have pointed out. That is:

$$\Sigma(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2 \Sigma \frac{1}{\lambda_i} > \beta'\beta + \sigma^2/\lambda_{min} \qquad (1.6)$$

Thus, when the degree of multicollinearity increases λ_{min} decreases; and the greater the degree of multicollinearity the longer the estimated coefficient vector $\hat{\beta}$.

It may be concluded from properties (1) and (2) that for an ill-conditioned data set there is a high probability that the estimated coefficient vector, $\hat{\beta}$, will be "off" not only in magnitude but also in direction. This in turn leads to the conclusion that the estimated coefficients will be overly sensitive to the sampling error or model misspecification due to the large variance.

Obviously, under these conditions the inflated β estimates of prediction equations will not accurately reflect the net relative importance of the variables under consideration. By the same token some non-significant coefficients may in fact be interpreted as important because they will be spuriously high. It is not unknown, for example, in situations where multicollinearity is a problem for the estimated coefficient to be larger than the correlation coefficient between the predictor and dependent variable. Indeed, occasionally, OLS standardized regression coefficients may be larger than one.

1.2.3 When an estimation data set is ill-conditioned OLS regression models do not provide accurate prediction even when the coefficient of determination (R^2) is high.

Though this is a well known fact among data analysts; and stems directly from the above discussion about regression coefficients being unstable and unreliable when estimated from ill-conditioned data, the fact need not necessarily be clearly understood. The conclusion may be formulated explicitly as follows.

The forecasting error variance, σ_f^2, of a multiple linear regression model, $y = X\beta + \varepsilon$, can be written as:

$$\sigma_f^2 = \sigma^2 \, [1 + \underset{\sim}{x}'V\underset{\sim}{x}] \tag{1.7}$$

where $\underset{\sim}{x}$ is a (p x 1) column vector of a single observation of the predictor variables, X's; V is the variance-covariance matrix of the estimated regression coefficients, which is $(X'X)^{-1}$ for OLS estimates, and σ^2 is the variance of the random error in y. Therefore, for OLS estimates forecasting error variance may be expressed as:

$$\begin{aligned} \sigma_f^2 &= \hat{\sigma}^2[1 + \underset{\sim}{x}(X'X)^{-1}\underset{\sim}{x}] \\ &= \hat{\sigma}^2[1 + \underset{\sim}{x}^{*}{}'\Lambda^{-1}\underset{\sim}{x}^{*}] \\ &= \hat{\sigma}^2[1 + \Sigma \, \frac{x_i^{*2}}{\lambda_i}] \end{aligned} \tag{1.8}$$

$$\text{or} \quad \hat{\sigma}_f^2 = \frac{\underline{y'y}(1-R^2)}{n-p} \, [1 + \Sigma \frac{x_i^{*2}}{\lambda_i}] > \frac{\underline{y'y}(1-R^2)}{n-p} \; \frac{x_p^{*2}}{\lambda_{min}} \tag{1.9}$$

where $\underset{\sim}{x}^* = P'\underset{\sim}{x}$ and P is the orthogonal transformation matrix of (X'X), such that $P'(X'X)P = \Lambda$, the eigenvalue matrix of (X'X), and $P'P = PP' = I$. From the above expression of forecasting error variance, it is clear in the first place that the forecasting error variance can be very large for ill-conditioned data for which λ_{min} is very small. Second, it is clear that the forecasting error variance is more sensitive to the smallest eigenvalue, λ_{min}, than to R^2. This is why in OLS regression, when the problem has a high degree of multicollinearity, high R^2 does not produce accurate prediction.

1.2.4 *The OLS estimates, $\hat{\beta}$'s, are sensitive to sampling fluctuation when the data set is ill-conditioned.*

From Eq. (1.5) it can be seen that for an ill-conditioned data set the OLS estimates have very large sampling variance due to their inverse relationship with the eigenvalues. In fact the sampling variance will approach infinity when the data set approaches perfect collinearity. Essentially, this implies that the OLS estimates are sensitive to sampling fluctuation.

1.2.5 *The estimated regression coefficients may be highly dependent and therefore error prone.*

In the discussion centering on Eq. (1.3) it was noted that when a data set has a high degree of multicollinearity, the estimates have large variance and covariance. This property accounts for the high correlations between estimates; and this is particularly true for highly correlated predictor variables. Thus, because of the high dependency between predictors, if one is erroneous, the other will also be erroneous. In general, when two predictors are highly positively correlated, the estimated coefficients are highly negatively correlated; that is, if one coefficient is over-estimated, the other will be under-estimated. This high correlation between estimates, plus the large variance of the estimates accounts for the fact that OLS estimates generally have too large a sum of squares

$(\hat{\beta}'\hat{\beta})$, and why they may even have the wrong sign.

1.2.6 *The OLS estimates are sensitive to model misspecification when the data set is ill-conditioned.*

It is well known that for orthogonal data sets, the regression coefficients ($\hat{\beta}$'s) are not affected by model misspecification due to the inclusion or exclusion of certain relevant variables. That is, the $\hat{\beta}$'s are invariant of the model specification. When the data set is ill-conditioned, however, the $\hat{\beta}$'s are highly correlated and consequently no longer invariant of model specification. In the presence of this condition the regression coefficients of included variables might change drastically as some relevant variables are dropped from, or added to the model. To what extent these coefficients change is largely dependent on the importance of the variable, and the extent to which it affects the variance inflation factor of each variable in the predictor variable set.

1.3 Measures of Multicollinearity

Obviously multicollinearity is an undesirable problem and one to be avoided if at all possible. As the above discussion indicates, however, in social science research and in the natural sciences as well, it is impossible to impose external controls over all the environmental constraints accounting for the correlations between conceptually distinct predictor variables (Marquardt & Snee, 1975). As Kish (1975, p. 266) points out, independence between social variables seldom exists; thus, if a sample is large enough the relationship between any two variables is nearly always greater than zero. Since multicollinearity is always a matter of degree and, therefore, least squares procedures always produce estimates that are inflated, overly sensitive to specification error, and possibly unreliable, the extent of the seriousness of the problem depends on the level of multicollinearity. Thus, with OLS regression the key question is: At what level of multicollinearity are the results unacceptable? This question in turn demands a measure of the severity of multicollinearity.

Several indicators of multicollinearity have been used by different researchers (e.g., Price, 1977; Farrar & Glauber, 1976; White, 1978). But Marquardt & Snee (1975) suggest that the maximum variance inflation factor is the best single measure. The variance inflation factors (VIFs) are the diagonal elements of the inverse of the simple correlation matrix. The VIF of each term is a measure of the collective impact of the interdependency of the explanatory variables on the variance of the regression coefficient of that term; and its value gives an indication of the number of times the variance has been inflated. When maximum VIF is used to measure the level of multicollinearity, one stands for no collinearity and infinity for perfect multicollinearity. The degree of multicollinearity can, however, be transformed into a scale of zero to one by defining the degree of multicollinearity "D" as:

$$D = D_{max} = \frac{2}{\pi}\tan^{-1}(V_{max} - 1) \qquad (1.10)$$

where V_{max} represents the maximum VIF. Hence, $D = 0$ for no collinearity, and $D = 1$ for perfect collinearity. Note that $\tan^{-1}(V_{max} - 1)$ is measured in radians.

It is too early in terms of this research tradition to present any rules of thumb. There are too many contingencies such as whether the research is exploratory, explanatory, or predictive; whether it is policy oriented research or discipline oriented and the like. Research conducted by the authors and reported elsewhere (see the monograph references) has been predominantly explanatory and discipline oriented. For these purposes "D" scale values greater than or equal to 0.7 have been considered serious, and D-values ≥ 0.5 have always been examined carefully. That is, where $D \geq 0.5$, OLS estimates have always been compared to those generated by a biased, reduced variance regression method. Collinearity of magnitude $D \geq 0.5$ is almost the norm in social science research where there are, say, four or more predictor variables; and collinearity of magnitude $D \geq 0.7$ is commonplace.

Obviously, the claim that $D \geq 0.7$ in explanatory, discipline-oriented models, is generally serious is rather subjective. Even for $D = 0.5$ (equivalent to $V_{max} = 2$) if one of the regression coefficients is small enough such that the inflated variance spans zero, the estimated OLS regression coefficients can be of the wrong sign; and in such an instance the effect of multicollinearity is definitely serious. When $D = 0.7$, the V_{max} is about 3 and the D_{max}/V_{max} relationship becomes steeper. Thus, in these instances D-values ≥ 0.7 can be considered serious regardless of the size of the regression coefficients. Otherwise, whether the D-value magnitude may be considered serious depends on the case.

It should be noted that the maximum effect of multicollinearity is manifest in factor space when the model is expressed in its canonical form. In view of this fact it may be preferred to measure the absolute degree of multicollinearity in factor space, for which $V_{max} = 1/\lambda_{min}$ and

$$D_{max} \leq D_c = \frac{2}{\pi} \tan^{-1}(1/\lambda_{min} - 1) \qquad (1.11)$$

where λ_{min} is the smallest eigenvalue.

The difference between D_c and D_{max} is that D_c is an absolute measure of the degree cf multicollinearity. It is a fixed value for a data set, which is invariant of a change in the data manipulation technique. It depends solely on the structure of the data set in the factor space. D_{max}, in contrast, is a relative measure in the sense that it depends on the orientation of the axes of the variable vectors. Its value changes with the data handling technique.

It is always true that $D_c \geq D_{max}$. The equals sign holds in the special case when one of the variable vectors lies on the major principal axis of the data set. The advantage of the V_{max} measure is that it is a linear scale while the D-measures are not. The D-measures, however, give a range of 0 to 1 for no collinearity to perfect collinearity, and is the same type of measure as the correlation coefficient. Thus, like the correlation coefficient the D_{max} measure may take negative values. Positive D_{max} signifies inflated variance ($V_{max} > 0$) while negative D_{max} signifies deflated or over-suppressed variance ($V_{max} < 0$).

1.4 Standard Solutions

It was noted above Eq. (1.4) that the total variance of the estimated regression coefficients based on the OLS procedure can be expressed as:

$$\Sigma Var(\hat{\beta}) = \sigma^2 Tr(X'X)^{-1}$$

$$= \frac{\varepsilon'\varepsilon}{n-p} Tr(X'X)^{-1} \qquad (1.12)$$

where ε is the random error vector of y in the standardized regression model, $y = X\beta + \varepsilon$, n is the sample size, p is the number of explanatory variables, and $Tr(X'X)^{-1}$ is the trace of the inverse of the correlation matrix (X'X). See Hoerl & Kennard (1970a).

Since in any statistical procedure, it is desirable to have variance as low as possible, and since from Eq. (1.12) it is obvious that the total variance depends on three independent factors (namely, $\varepsilon'\varepsilon$, n-p, and $Tr(X'X)^{-1}$, then standard solutions to variance inflation depend on the location of its sources followed by efforts to eliminate them if possible. These sources are:

1.4.1 The random error. This is the purely random portion of the dependent variable y. It generally consists of two parts; that due to measurement error ε_m, and that due to stochastic error ε_s. Stochastic error is that which is attributable to two components in turn, (i) the incompleteness of the designed model, and (ii) some unknown irreproducible fluctuation (cf. Wannacott & Wannacott, 1969, p. 17). Since these two errors are assumed to be uncorrelated and to have normal distribution with zero mean, the sum of squares of random error $\varepsilon'\varepsilon$ can be expressed as:

$$\varepsilon'\varepsilon = \varepsilon_m'\varepsilon_m + \varepsilon_s'\varepsilon_s$$

This random error can be reduced but not eliminated (i) by improving the precision of measurement, and (ii) by improving the designed model; that is, by improved model specification.

1.4.2 The sample size, n. More specifically, this is the degree of freedom of the sum of squares error (n-p). Thus total variance can be reduced, but again not eliminated, by increasing the sample size.

From sources (1) and (2) we have derived the three best known standard solutions. For multiple linear regression the total variance of the estimated coefficients may be reduced by improving the measurement of the variables, by improving the specification of the model, and by increasing the sample size.

But the most important source of variance inflation is that due to multicollinearity. This matter is referred to next, but a more detailed analysis of the standard statistical solution to the multicollinearity problem is presented in the next chapter.

1.4.3 The degree of multicollinearity. The effect of the degree of multicollinearity enters through $Tr(X'X)^{-1}$, which is the trace of $(X'X)^{-1}$; that is, the sum of the variance inflation factor of each variable due to its interrelationship with the rest of the explanatory variables. It is noteworthy that this third source of variance inflation, unlike the other two, is statistical. Thus, a statistical solution is called for.

1.5 Summary and Overview

In this chapter an effort was made

(1) to introduce the concept of multicollinearity;

(2) to explain why multicollinearity is a problem for the analyst working in the structural equation tradition;

(3) to show that, though the problem is a general one in all the sciences, it is particularly serious and extensive in the social sciences;

(4) to identify and explain the six harmful effects of multicollinearity for least squares estimation;

(5) to describe three useful measures of multicollinearity - the maximum variance inflation factor (V_{max}), the maximum degree of multicollinearity (D_{max}), and an absolute degree of multicollinearity (D_c) which is based on a model expressed in its canonical form;

(6) to suggest ways of determining the seriousness of the multicollinearity problem based on V_{max}, D_{max} and D_c indices without invoking simplistic rules of thumb; and

(7) to describe the sources of variance inflation in the regression coefficients in order to identify both the

standard non-statistical methods of coping with the problem as well as the statistical solutions.

It was noted that the most important solution is a statistical one. Thus, the remaining chapters in the monograph deal with the theoretical and applied aspects of these statistical solutions. The ridge regression family of methods are described in theory and practice in chapters two and three. The ridge method is well established in the Technometrics literature. Nevertheless, some of the problems attendant upon the use of ridge regression procedures are identified and in chapter four a procedure developed by the authors called simple normalization regression is presented because it seems to cope satisfactorily with some of the inherent limitations of ridge regression for the estimation of stochastic, misspecified, social science models.

CHAPTER TWO

2.0 SIMPLE RIDGE REGRESSION IN THEORY

2.1 Simple Ridge Regression

As was pointed out in chapter one, ordinary least squares regression based on a minimum residual sum of squares criterion does not produce satisfactory results, because the acceptability of OLS estimates depends on the degree of multicollinearity of the data set; and the acceptability worsens as the degree of multicollinearity increases. Hoerl & Kennard (1970a), however, have developed a promising rival technique called "ridge regression", which reduces multicollinearity. It is based on a minimum total mean square error criterion as opposed to the minimum residual sum of squares criterion.

In ridge regression, the statistical model and its assumptions are the same as those for OLS regression which were presented in chapter one (section 1.2). One minor difference is that in ridge regression estimation all the variables (the X's and y) are standardized in order that X'X gives the correlation matrix of the predictors and X'y gives the vector of correlation coefficients of the dependent and independent variables.

In Hoerl & Kennard's ridge regression the criterion for measuring the goodness of an estimator $\hat{\beta}^*$ is the total mean square error (MSE) function defined by:

$$\text{MSE} = E[(\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta)] \qquad (2.1)$$

MSE is the sum of the mean square error of the individual regression coefficients, and it can be proven (Pindyck & Rubinfeld, 1976, p. 22, or see property (vii) of ridge regression in this chapter) that it can be decomposed into the sum of the total variance and the total squared bias of each regression coefficient; that is,

$$MSE = \Sigma Var(\hat{\beta}_i^*) + Bias^2(\hat{\beta}^*) \qquad (2.2)$$

From the geometrical point of view, MSE represents the expected squared distance between $\hat{\beta}^*$ and the true vector β; thus, a "good" estimator would be the one that minimizes this distance - that is, minimizes the MSE. Further, since the least squares estimator, $\hat{\beta}$, gives the minimum residual sum of squares; $\hat{\beta}^*$, the estimator based on the minimum MSE criterion, will give an inflated residual sum of squares, which can be written as:

$$\begin{aligned} \phi = \varepsilon'\varepsilon &= (y - X\hat{\beta}^*)'(y - X\hat{\beta}^*) \\ &= (y - X\hat{\beta})'(y - X\hat{\beta}) + (\hat{\beta}^* - \hat{\beta})'X'X(\hat{\beta}^* - \hat{\beta}) \\ &= \phi_{min} + \phi_o(\hat{\beta}^*) \qquad (2.3) \end{aligned}$$

where

$$\phi_o(\hat{\beta}^*) = (\hat{\beta}^* - \hat{\beta})'X'X(\hat{\beta}^* - \hat{\beta}) \qquad (2.4)$$

is the inflation in the residual sum of squares. Geometrically, it gives the surface of a family of hyperellipsoids centered at $\hat{\beta}$, which is the OLS estimate of β.

The so-called ridge estimator, $\hat{\beta}^*$, is the one which minimizes the squared length of the coefficient vector subjected to constant inflation in the residual sum of squares, $\phi_o(\hat{\beta}^*)$, which is determined by the minimum MSE. That is, $\hat{\beta}^*$ is a solution to minimizing $\hat{\beta}^{*\prime}\hat{\beta}^*$ subjected to

$$(\hat{\beta}^* - \hat{\beta})'X'X(\hat{\beta}^* - \hat{\beta}) = \phi_o \text{ (a constant determined by the minimum MSE).}$$

The solution of this problem is obtained by minimizing the Lagrangian function

$$f(\hat{\beta}^*) = \hat{\beta}^{*\prime}\hat{\beta}^* + \frac{1}{k}[(\hat{\beta}^* - \hat{\beta})'X'X(\hat{\beta}^* - \hat{\beta}) - \phi_o] \qquad (2.5)$$

where $\frac{1}{k}$ is the Lagrangian multiplier. At minimum $f(\hat{\beta}^*)$, we have

$$\frac{\partial}{\partial\hat{\beta}^*} f(\hat{\beta}^*) = 2\hat{\beta}^* + \frac{1}{k}\,2X'X(\hat{\beta}^* - \hat{\beta}) = 0$$

$$(X'X + kI)\hat{\beta}^* = X'X\hat{\beta} = X'y$$

therefore,

$$\hat{\beta}^* = (X'X + kI)^{-1}\,X'y \qquad (2.6)$$

where k is a parameter determined by Eq. (2.4) which is in turn determined by minimum MSE (see Hoerl & Kennard, 1970a). It will be pointed out below in the discussion of the optimal k that because of a cyclical relationship between k, ϕ_o, and $\hat{\beta}^*$, MSE alone cannot complete the ridge procedure. (See especially, property 6, property 7, and property 8 of simple ridge regression described in this chapter.) Either k or ϕ_o has to be predetermined by a different method, which means in practice that numerous methods can be devised for determining k; unfortunately, none can guarantee to give the true k which minimizes MSE.

Graphically, the geometrical relation of the ridge regression procedure in two dimensional space can be depicted by Figure 2.1; where the ellipse is the hyperellipsoid of $\phi_o(\hat{\beta}^*) = \phi_o$ = a constant; and where the ridge estimator $\hat{\beta}^*$ is a vector which is the shortest among those that have constant inflated residual sum of squares ϕ_o.

Figure 2.1 About Here

From the above argument we have seen that the minimum square length of the coefficient vector, or the minimum sum of squared regression coefficients determines the form of the ridge estimator, and the minimum MSE determines the value of the parameter k. Note that the form of the ridge estimator, $\hat{\beta}^*$ in Eq. (2.6) is the same as that derived by Lindly & Smith (1972) using Bayesian methods under the assumption of an exchangeable prior distribution for β.

2.2 The Properties of Simple Ridge Regression

Though most of the properties of ridge regression have been thoroughly discussed by researchers such as Hoerl & Kennard (1970a), and Marquardt (1970), they are brought together here. Some sixteen properties are summarized. In several places the observations of others have been extended.

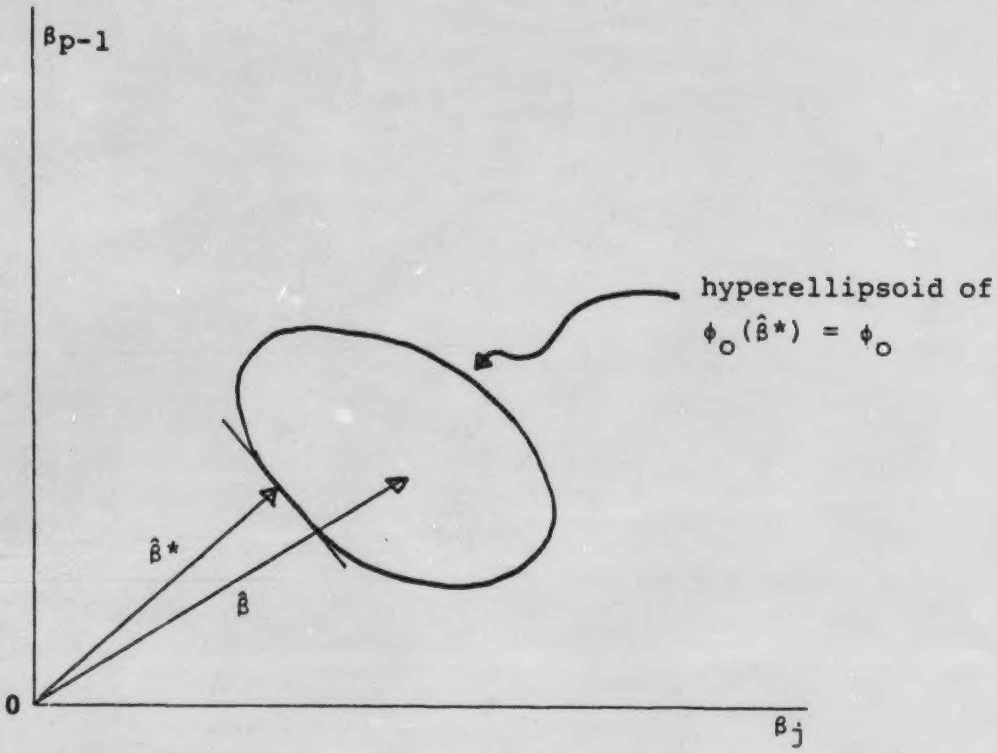
2.2.1 *Ridge regression gives a shorter regression coefficient vector than that of OLS regression.*

Proof: For OLS regression, since $\hat{\beta} = (X'X)^{-1}X'y$, we have

$$\hat{\beta}'\hat{\beta} = y'X(X'X)^{-2}X'y = y'XP\Lambda^{-2}P'X'y$$

$$= y'X^*\Lambda^{-2}X^{*\prime}y = \Sigma \frac{(X^{*\prime}y)^2{}_i}{\lambda_i} \qquad (2.7)$$

Figure 2.1 The Geometrical Relationship of the OLS Estimate, the Ridge Regression Estimate, and the Inflation in the Residual Sum of Squares.


β_{p-1}
hyperellipsoid of
$\phi_o(\hat{\beta}^*) = \phi_o$
$\hat{\beta}^*$
$\hat{\beta}$
0
β_j

where $X^* = XP$ and P is the eigenvector matrix of $X'X$ which is an orthogonal transformation matrix satisfying $P'P = PP' = I$, and Λ is the eigenvector matrix of $X'X$ with eigenvalues λ_1, λ_2, ..., λ_ϕ, satisfying $X'X = P\Lambda P'$.

For ridge regression, $\hat{\beta}^* = (X'X + kI)^{-1}X'y$, and

$$\begin{aligned}\hat{\beta}^{*\prime}\hat{\beta}^* &= y'X(X'X + kI)^{-2}X'y \\ &= y'X^*(\Lambda + kI)^{-2}X^{*\prime}y \\ &= \Sigma \frac{(X^{*\prime}y)^2{}_i}{(\lambda_i + k)^2} \qquad k \geq 0 \qquad (2.8)\end{aligned}$$

From the above two relations (cf. Eqs. 2.7 and 2.8), it is obvious that, $\hat{\beta}^{*\prime}\hat{\beta}^* \leq \hat{\beta}'\hat{\beta}$; hence, ridge regression ($k > 0$) gives a shorter regression coefficient vector.

2.2.2 The ridge estimator $\hat{\beta}^*$ is a linear transformation of the OLS estimator $\hat{\beta}$, which is

$$\hat{\beta}^* = Z\hat{\beta} \qquad (2.9)$$

where

$$Z = (X'X + kI)^{-1}X'X, \text{ or } Z = P \text{ Diag}\left[\lambda_i/(\lambda_i + k)\right] P',$$

when it is expressed in terms of the eigenvalues of $X'X$, and where Diag $[\lambda_i/(\lambda_i + k)]$ represents a diagonal matrix with ith diagonal element $\lambda_i/(\lambda_i + k)$.

Proof: The ridge estimator

$$\begin{aligned}\hat{\beta}^* &= (X'X + kI)^{-1}X'y \\ &= (X'X + kI)^{-1}X'(X\hat{\beta} + \hat{\varepsilon}) \\ &= (X'X + kI)^{-1}X'X\hat{\beta} \qquad (2.10)\end{aligned}$$

where use is made of the fact that in OLS regression the residual vector is perpendicular to all independent variables

X's; that is, $X'\hat{\varepsilon} = 0$. Therefore, the linear transformation is

$$Z = (X'X + kI)^{-1}X'X \qquad (2.11)$$

and when expressed in terms of the eigenvalues of X'X, we have

$$\begin{aligned} Z &= P(\Lambda + kI)^{-1}\,\Lambda P' \\ &= P\ \mathrm{Diag}\,[\lambda_i/(\lambda_i + k)]P' \qquad (2.12) \end{aligned}$$

where

$$\mathrm{Diag}\ \lambda_i/(\lambda_i + k) = \begin{array}{cccc} \lambda_i/(\lambda_i + k) & & & \\ & \lambda_2/(\lambda_2 + k) & & \\ & & \ddots & \\ & & & \lambda_p/(\lambda_p + k) \end{array}$$

2.2.3 *The ridge estimator $\hat{\beta}^*$ has a variance-covariance matrix*

$$\mathrm{cov}(\hat{\beta}^*) = \sigma^2[\mathrm{VIF}] \qquad (2.13)$$

where [VIF] is the variance inflation factor matrix, and

$$[\mathrm{VIF}] = (X'X + kI)^{-1}X'X(X'X + kI)^{-1} \qquad (2.14)$$

or, when expressed in terms of the eigenvalue of the correlation matrix, X'X,

$$[\mathrm{VIF}] = P\ \mathrm{Diag}[\lambda_i/(\lambda_i + k)^2]P' \qquad (2.15)$$

where Diag $[\lambda_i/(\lambda_i + k)^2]$ is a diagonal matrix with $\lambda_i/(\lambda_i + k)^2$ as its ith element.

Proof: Since the ridge estimator is a linear transformation of $\hat{\beta}$, i.e., $\hat{\beta}^* = Z\hat{\beta}$, we have

$$\begin{aligned} \mathrm{cov}(\hat{\beta}^*) &= \mathrm{cov}(Z\hat{\beta}) \\ &= Z\mathrm{cov}(\hat{\beta})Z \\ &= \sigma^2 Z(X'X)^{-1}Z' \\ &= \sigma^2(X'X + kI)^{-1}X'X(X'X + kI)^{-1} \end{aligned}$$

Therefore,

$$[VIF] = (X'X + kI)^{-1}X'X(X'X + kI)^{-1}$$

When it is expressed in terms of the eigenvalues of X'X, we apply $X'X = P\Lambda P'$, and obtain

$$\begin{aligned} [VIF] &= P(\Lambda + kI)^{-1} \Lambda(\Lambda + kI)^{-1} P' \\ &= P \, \mathrm{Diag}[\lambda_i(\lambda_i + k)^2]P' \end{aligned}$$

2.2.4 *Ridge regression produces smaller variances of the regression coefficients than that of OLS regression; however it does not necessarily reduce the covariance or correlation between them.*

Proof: The variance-covariance matrix of ridge estimates is

$$\mathrm{cov}(\hat{\beta}^*) = \sigma^2[VIF]$$

where

$$[VIF] = P \, \mathrm{Diag}\,[\lambda_i/(\lambda_i + k)^2]P'$$

is the variance-covariance inflation factor matrix with elements

$$VIF_{jl} = \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} p_{ji}p_{li} \quad \text{for all } j,\ l = 1, 2, \ldots, p$$

and where p_{ji}, which can be positive or negative, represents the ith element of the jth eigenvalue in orthogonal transformation matrix P. Since $k > 0$ for ridge regression and $k = 0$ for OLS regression, the variances

$$VIF_{jj} = \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} p_{ji}^2 \quad \text{for all } j = 1, 2, \ldots, p \tag{2.16}$$

are reduced by the ridge procedure. However, the covariances

$$VIF_{jl} = \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} p_{ji}p_{li} \quad \text{for } j \neq l \text{ and all } j, \; j, l = 1, 2, \ldots, p \tag{2.17}$$

can be inflated or deflated due to the uncertainty in the sign of "$p_{ji}p_{li}$", as compared to that of OLS regression.

From the above, it is clear, that the correlation between $\hat{\beta}_j^*$ and $\hat{\beta}_l^*$, which is

$$r_{jl}^* = \frac{VIF_{jl}}{\sqrt{VIF_{jj}} \; \sqrt{VIF_{ll}}}$$

can also be inflated or deflated.

Comments: 1. When a simple ridge procedure is used, its effects on the correlation between the regression coefficients should be checked and the analyst should interpret the estimated coefficients with caution. If the correlation coefficients are significantly different from zero, the statistical control of the variables will not have been achieved.

2. In a generalized ridge regression, therefore, one should look not only at the variance but also at the covariance in order to reduce the correlation coefficients between the regression coefficients.

2.2.5 *The total variance of ridge estimates*

$$\Sigma \mathrm{Var}(\hat{\beta}_i^*) = \sigma^2 \Sigma \lambda_i/(\lambda_i + k)^2 \qquad (2.18)$$

is a monotonic decreasing function of k.

Proof: The variance-covariance matrix (property 3) of ridge estimates is

$$\mathrm{cov}(\hat{\beta}^*) = \sigma^2 P \, \mathrm{Diag}[\lambda_i/(\lambda_i + k)^2] P'$$

Therefore, the total variance of the estimates is

$$\begin{aligned} \Sigma \mathrm{Var}(\hat{\beta}_i^*) &= \mathrm{Tr}\, \mathrm{cov}(\hat{\beta}^*) \\ &= \sigma^2 \mathrm{Tr}\, P \, \mathrm{Diag}\, [\lambda_i/(\lambda_i + k)^2]\, P' \\ &= \sigma\, \mathrm{Tr}\, \mathrm{Diag}\, [\lambda_i/(\lambda_i + k)^2]\, P'P \\ &= \sigma^2 \Sigma \lambda_i/(\lambda_i + k)^2 \end{aligned}$$

It is obvious then, since k and all λ_i for $i = 1, 2, \ldots, p$, are positive values, the total variance is a monotonic decreasing function of k.

Comments: 1. The total variance has a range of $(\sigma^2 \Sigma 1/\lambda_i, 0)$ when k varies from 0 to infinity. This means that the total variance reduces to zero when k approaches infinity. However, at large k, the variance reduces at a much slower rate.

2. The decreasing rate at which $k \to 0$, that is,

$$\lim_{k \to 0} \left| \frac{d}{dk} \Sigma \mathrm{Var}(\hat{\beta}_i^*) \right| = 2\sigma^2 \Sigma \frac{1}{\lambda_i^2} > \frac{2\sigma^2}{\lambda_{\min}} \qquad (2.19)$$

can be very large for problems with a high degree of multicollinearity. This implies that ridge regression is more effective for problems with a high degree of multicollinearity.

2.2.6 *The ridge estimator $\hat{\beta}^*$ is a negatively biased estimator, with a bias given by*

$$\text{Bias}(\hat{\beta}^*) = -k(X'X + kI)^{-1}\beta \qquad (2.20)$$

or

$$\text{Bias}(\hat{\beta}^*) = k\ P\ \text{Diag}[1/(\lambda_i + k)]P'\beta \qquad (2.21)$$

when it is expressed in terms of the eigenvalues of the correlation matrix.

Proof:

$$\begin{aligned} \text{Bias}(\hat{\beta}^*) &= E(\hat{\beta}^*) - \beta = Z\beta - \beta \\ &= [(X'X + kI)^{-1}X'X - I]\beta \\ &= -k(X'X + kI)^{-1}\beta \end{aligned}$$

In terms of the eigenvalues, by substituting, $X'X = P\Lambda P'$ and $P'P = I$, we have

$$\text{Bias}(\hat{\beta}^*) = -k\ P\ \text{Diag}[1/(\lambda_i + k)]P'\beta$$

Comment: 1. It is obvious that the bias produced by ridge regression is a function of an unknown population regression coefficient vector β, and hence the bias cannot be calculated.

2. Since $\hat{\beta}^*$ is negatively biased, if we are able to prove $\hat{\beta}^*$ or $E(\hat{\beta}^*)$ is significant, then the true value β, must be significant.

2.2.7 *The total square bias of ridge estimates*

$$\text{Bias}^2(\hat{\beta}^*) = k^2\Sigma\frac{\alpha_i}{(\lambda_i + k)^2} \qquad (2.22)$$

is a monotonic increasing function of k with a range of $(0,\Sigma\beta_i^2)$ for k = 0 to infinity.

Proof: From property (6) we have the square bias of the estimates

$$\begin{aligned}\text{Bias}^2(\hat{\beta}^*) &= k^2\beta'P\,\text{Diag}\,[1/(\lambda_i + k)]\,\text{Diag}\,[1/(\lambda_i + k)]\,P'\beta \\ &= k^2\alpha'\text{Diag}\,[1/(\lambda_i + k)^2]\,\alpha \\ &= k^2\,\Sigma\,\frac{\alpha_i^2}{(\lambda_i + k)^2}\end{aligned}$$

where $\alpha = P'\beta$ is the true regression coefficient vector in canonical form $y = X^*\alpha + \varepsilon$, and where $X^* = XP$. It is obvious, therefore, that $\text{Bias}^2(\hat{\beta}^*)$ is a monotonic increasing function of k. Further, since

$$\lim_{k\to 0} \text{Bias}^2(\hat{\beta}^*) = \lim_{k\to 0} k^2\Sigma\frac{\alpha_i^2}{(\lambda_i + k)^2} = 0$$

and

$$\begin{aligned}\lim_{k\to\infty} \text{Bias}^2(\hat{\beta}^*) &= \lim_{k\to\infty} k^2\Sigma\frac{\alpha_i^2}{(\lambda_i + k)^2} \\ &= \lim_{k\to\infty} \Sigma\frac{\alpha_i^2}{(\lambda_i/k + 1)^2} \\ &= \Sigma\alpha_i^2 \\ &= \Sigma\beta_i^2\end{aligned}$$

the total square bias of ridge estimates has a range $(0, \Sigma\beta_i^2)$.

Comment: The total square bias depends on the unknown population regression coefficient vector β; therefore, it can never be obtained.

2.2.8 *The ridge estimator $\hat{\beta}^*$ has a total mean square error (MSE) of*

$$MSE = \sigma^2 \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \Sigma \frac{\alpha_i^2}{(\lambda_i + k)^2} \qquad (2.23)$$

Proof: The total mean square error

$$MSE = E(\hat{\beta}^* - \hat{\beta})^2$$

$$= E\{[\hat{\beta}^* - E(\hat{\beta}^*)] + [E(\hat{\beta}^*) - \beta]\}^2$$

$$= E[\hat{\beta}^* - E(\hat{\beta}^*)]^2 + [E(\hat{\beta}^*) - \beta]^2$$

$$= \text{Tr cov } (\hat{\beta}^*) + \text{Bias}^2(\hat{\beta}^*)$$

$$= \sigma^2 \text{Tr P Diag}[\lambda_i/(\lambda_i + k)^2]P' + k^2\alpha' \text{ Diag } [1/(\lambda_i + k^2)]\alpha$$

$$= \sigma^2 \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \Sigma \frac{\alpha_i^2}{(\lambda_i + k)^2}$$

Comments: 1. From this relation, it is obvious that the total mean square error of ridge regression depends on the true unknown parameters σ^2 and α's. Therefore, the total mean square error cannot be obtained.

2. The first component of MSE is the total variance of estimation; that is,

$$\Sigma \text{Var}(\hat{\beta}_i^*) = \sigma^2 \Sigma \frac{\lambda_i}{(\lambda_i + k)^2}$$

which describes the random portion of the error, while the second component is the square bias; that is,

$$\text{Bias}^2(\hat{\beta}^*) = k^2 \Sigma \frac{\alpha_i}{(\lambda_i + k)^2}$$

which is the systematic portion of the error.

2.2.9 Ridge regression gives a minimum distance between $\hat{\beta}^*$ and the true vector β, which in this sense makes $\hat{\beta}^*$ a better estimator than that of $\hat{\beta}$, the OLS estimator.

Proof: The ridge regression criterion demands an estimation procedure minimizing the mean square error. From a geometrical point of view MSE is the squared distance from $\hat{\beta}^*$ to β; therefore, the ridge estimator gives a minimum distance from $\hat{\beta}^*$ to the true vector β.

Comment: Since the MSE of ridge regression depends on the unknown parameters σ^2 and α, the minimum MSE or the minimum distance from $\hat{\beta}^*$ to β cannot be obtained.

2.2.10 *Ridge regression inflates the residual sum of squares of*

$$\phi_o = k^2\hat{\beta}^{*\prime}(X'X)^{-1}\hat{\beta}^* \tag{2.24}$$

or

$$\phi_o = k^2\Sigma\hat{\alpha}_i^{*2}/\lambda_i \tag{2.25}$$

where $\hat{\alpha}^* = P'\hat{\beta}^*$ is the ridge regression coefficient vector in a factor space defined by orthogonal transformation $X^* = XP$, in which the multiple linear regression model in canonical form is given by $y = X^*\alpha + \varepsilon$.

Proof: Since

$\hat{\beta}^* = Z\hat{\beta}$ or $\hat{\beta} = Z^{-1}\hat{\beta}$ we have

$$\begin{aligned}\hat{\beta}^* - \hat{\beta} &= (I - Z^{-1})\hat{\beta}^* \\ &= [I - (X'X)^{-1}\ (X'X + kI)]\,\hat{\beta}^* \\ &= -k\ (X'X)^{-1}\hat{\beta}^*\end{aligned}$$

Therefore

$$\begin{aligned}\phi_o &= (\hat{\beta}^* - \hat{\beta})'\ X'X(\hat{\beta}^* - \hat{\beta}) \qquad \text{Eq. (2.4)} \\ &= k^2\hat{\beta}^{*\prime}\ (X'X)^{-1}\hat{\beta}^*\end{aligned}$$

In terms of eigenvalues, by substituting $X'X = P\Lambda P'$ and $P'\hat{\beta}^* = \hat{\alpha}^*$ we have

$$\phi_o = k^2\hat{\alpha}^{*\prime}\Lambda^{-1}\hat{\alpha}^* = k^2\Sigma\hat{\alpha}_i^{*2}/\lambda_i$$

2.2.11 Ridge regression produces a smaller multiple R square than that of OLS regression, which can be expressed as

$$R^2 = \hat{\beta}^{*\prime}X'X\hat{\beta}^* \tag{2.26}$$

$$= \hat{\beta}^{*\prime}X'y - k\hat{\beta}^{*\prime}\hat{\beta}^* \tag{2.27}$$

or, when expressed in terms of the eigenvalues of X'X, as

$$R^2 = \Sigma\lambda_i\hat{\alpha}_i^{*2} \tag{2.28}$$

Proof: The Multiple R square

$$R^2 = \frac{RSS}{TSS} = \frac{\hat{y}^{*\prime}\hat{y}^*}{y'y} = \frac{\hat{\beta}^{*\prime}X'X\hat{\beta}^*}{y'y} = \hat{\beta}^{*\prime}X'X\hat{\beta}^*$$

where we have used $y'y = 1$, since in ridge regression, all variables are standardized to give unit length. Further, by substituting $\hat{\beta}^* = (X'X + kI)^{-1}X'y$ and using a little algebra, we have

$$\begin{aligned} R^2 &= \hat{\beta}^{*\prime}X'X\hat{\beta}^* \\ &= \hat{\beta}^{*\prime}X'X(X'X + kI)^{-1}X'y \\ &= \hat{\beta}^{*\prime}[(X'X + kI) - kI(X'X + kI)^{-1}X'y \\ &= \hat{\beta}^{*\prime}X'y - k\hat{\beta}^{*\prime}\hat{\beta}^* \end{aligned}$$

And by using $X'X = P\Lambda P'$ and $P'\hat{\beta}^* = \hat{\alpha}^*$ we obtain the expression in terms of the eigenvalues of X'X, that is,

$$R^2 = \hat{\beta}^{*\prime}X'X\hat{\beta}^* = \hat{\alpha}^{*\prime}\Lambda\hat{\alpha}^* = \Sigma\lambda_i\hat{\alpha}_i^{*2}$$

From Eq. (2.27), it can be seen that for OLS regression the multiple R^2 is greater than that of ridge regression, since for OLS regression we have $k = 0$ and $|\hat{\beta}| > |\hat{\beta}^*|$.

Comment: The R^2 for ridge regression is a monotonic decreasing function of k. This can be seen from the fact R^2 can be expressed as

$$R^2 = \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} (X^{*\prime}y)_i^2 \qquad (2.29)$$

which is obtained from substituting $\hat{\alpha}_i^* = (X^{*\prime}y)_i/(\lambda_i + k)$ into Eq. (2.28).

2.2.12 *The ridge estimate $\hat{\beta}^*$ is less sensitive to sampling fluctuation as compared to the OLS estimate.*

From property (4) we know that the ridge procedure produces smaller sampling variance. Essentially this implies that the ridge estimate is less sensitive to sampling fluctuation.

2.2.13 *The ridge estimate produces a more accurate prediction equation than the OLS regression procedure, if the bias introduced is not too large.*

Proof: It is well known that for unbiased estimates the variance of the forecasting error is

$$\sigma_f^2 = \sigma^2[1 + \underset{\sim}{x}'V\underset{\sim}{x}]$$

where $\sigma^2 V$ is the variance-covariance matrix of the estimated parameters. For a biased estimate, such as a ridge estimate, the square bias should be added to the forecasting error variance; that is,

$$\begin{aligned} \sigma^{2*}_f &= \sigma_f^2 + \text{Bias}^2(\hat{\beta}^*) \\ &= \sigma^2[1 + \underset{\sim}{x}'V\underset{\sim}{x}] + \text{Bias}^2(\hat{\beta}^*) \end{aligned}$$

Therefore, when it is expressed explicitly, we have

$$\begin{aligned}\sigma_f^{*2} &= \sigma^2\{1 + \underset{\sim}{x}'P\ \mathrm{Diag}[\lambda_i/(\lambda_i + k)^2]P'\underset{\sim}{x}\} + \mathrm{Bias}^2(\hat{\beta}^*) \\ &= \sigma^2\{1 + \underset{\sim}{x}^{*\prime}\mathrm{Diag}[\lambda_i/(\lambda_i + k)^2]\underset{\sim}{x}^*\} + \mathrm{Bias}^2(\hat{\beta}^*) \\ &= \sigma^2\left[1 + \Sigma\frac{\lambda_i x_i^{*2}}{(\lambda_i + k)^2}\right] + \mathrm{Bias}^2(\hat{\beta}^*) \qquad (2.33)\end{aligned}$$

For OLS estimates, we set k = 0 and derive the forecasting error variance as

$$\sigma_f^2 = \sigma^2\left[1 + \Sigma\frac{x_i^{*2}}{\lambda_i}\right] \qquad (2.34)$$

which is much larger than σ_f^{*2} if the bias produced by the ridge procedure is not too large relative to the reduction in variance.

Comments: 1. The forecasting error variance σ_f^{*2} consists of two parts. The first part is the variance term which describes the random part of the error, and the second part (the bias term) describes the systematic part of the error.

2. In the case where the bias term is relatively large, the more accurate prediction can still be obtained by dividing the sample (if large enough) into two sets, one is used to estimate the biased parameter, and the other is used to estimate the bias in the prediction of the dependent variable. The accuracy of this empirically estimated bias in prediction can be improved by repeating the procedure using different ways of dividing the data set.

2.2.14 *There exists a wide range of k, $0 < k > k_{max}$, which will give a set of ridge estimates $\hat{\beta}^*$; and which will produce smaller MSE than that of OLS estimates.*

Proof: Define the effectiveness index (Eft) of ridge regression as the ratio of reduction in total variance to the total square bias introduced by the ridge regression; that is,

$$\text{Eft} = \frac{\text{Reduction in total variance}}{\text{Bias}^2\ (\hat{\beta}^*)} \tag{2.35}$$

$$= \frac{\sigma^2\ \text{Tr}\ (X'X)^{-1} - \sigma^2\ \text{Tr}[\text{VIF}]}{\text{Bias}^2\ (\hat{\beta}^*)}$$

$$= \frac{\sigma^2 \left[\Sigma \frac{1}{\lambda_i} - \Sigma \frac{\lambda_i}{(\lambda_i + k)} \right]}{k^2\ \Sigma \frac{\sigma_i^2}{(\lambda_i + k)^2}} \tag{2.36}$$

Since the total variance of ridge regression is a monotonic decreasing function and $\text{Bias}^2(\hat{\beta}^*)$ is a monotonic increasing function of k, then Eft is a decreasing function of k. It can be proven that the effectiveness of ridge regression has a range of infinity to zero when k varies from 0 to infinity. Further, we have

$$\begin{aligned} \text{MSE}(\hat{\beta}) - \text{MSE}(\hat{\beta}^*) &= \Sigma\text{Var}(\hat{\beta}) - \Sigma\text{Var}(\hat{\beta}^*) - \text{Bias}^2(\hat{\beta}^*) \\ &= \text{Eft} \times \text{Bias}^2(\hat{\beta}) - \text{Bias}^2(\hat{\beta}^*) \\ &= (\text{Eft} - 1)\ \text{Bias}^2(\hat{\beta}^*) \end{aligned}$$

then, for any k which gives Eft > 1, we have

$$\text{MSE}(\hat{\beta}) - \text{MSE}(\hat{\beta}^*) > 0$$

That is, the ridge regression procedure produces smaller MSE. If we get $k = k_{max}$ (maximum k) for Eft(k) = 1, then all k's that are less than k_{max} would give smaller MSE.

Comments: 1. Any valid ridge procedure should produce an optimal k which is less than k_{max}.

2. The maximum k defined here is a function of unknown parameters σ^2 and α's, and hence the true value will never be obtained. However, if we use the OLS estimate of σ^2 and α's, we would obtain a conservative estimate of k_{max} (called $\hat{k}_{max}$), due to the fact that the α's are generally over-estimated by OLS regression.

3. Since $\hat{k}_{max}$ is a conservative estimate, $k < \hat{k}_{max}$ is a sufficient but not necessary condition for a valid ridge procedure.

4. $\hat{k}_{max}$ and the maximum k (called $k_{\check{max}}$) defined by Vinod (1976, 1978) refer to the same theoretical maximum k; however, empirically $\hat{k}_{max} > k_{\check{max}}$. That is, $\hat{k}_{max}$ is a more accurate estimate of $k_{\check{max}}$.

5. The OLS estimate of the effectiveness index

$$\frac{\hat{\sigma}^2 \left[\Sigma \frac{1}{\lambda_i} - \Sigma \frac{\lambda_i}{(\lambda_i + k)^2}\right]}{k^2 \, \Sigma \frac{\hat{\alpha}_i^2}{(\lambda_i + k)^2}}$$

may be used to indicate the performance of a ridge procedure. If $\hat{Eft}(k) > 1$, the ridge procedure is a valid one. However, as in comment (2), $\hat{Eft}(k)$ is a conservative estimate of the true effectiveness index, and therefore $\hat{Eft}(k) > 1$ is a sufficient not a necessary condition.

2.2.15 *For any problem there exists a positive "optimal k" (called k_o) which gives a minimum MSE.*

Proof: (Following Kasarda & Shih, 1977). Since

$$MSE = \Sigma Var(\hat{\beta}_i^*) + Bias^2(\hat{\beta}^*)$$

$$\frac{d}{dk} MSE = \frac{d}{dk} \Sigma Var(\hat{\beta}_i^*) + \frac{d}{dk} Bias^2(\hat{\beta}^*)$$

$$= -2\sigma^2 \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} + 2k \Sigma \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3}$$

$$= -2 \Sigma \frac{\lambda_i}{(\lambda_i + k)^3} (\sigma^2 - k\alpha_i^2 \qquad (2.39)$$

Based on Rolle's Theorem (Widder 1963), since

$$\lim_{k \to 0} \frac{d}{dk} MSE = -2\sigma^2 \Sigma \frac{1}{\lambda_i^2} < 0$$

the optimal k is a positive value. Further, since

$$\frac{d^2}{dk^2} MSE = 6\sigma^2 \Sigma \frac{\lambda_i}{(\lambda_i + k)^4} + 2\Sigma \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^4} (\lambda_i - 2k) \qquad (2.40)$$

and

$$\lim_{k \to 0} \frac{d^2}{dk^2} MSE = 6\sigma^2 \Sigma \frac{1}{\lambda_i^3} + 2\Sigma \frac{\alpha_i}{\lambda_i} > 0 \qquad (2.41)$$

According to the theory of minimum (Widder 1963), this positive value leads to the conclusion that a minimum exists for MSE (Kasarda and Shih, 1977).

Comment: This is what Hoerl & Kennard (1970a) called the "existence theorem". It is true even for an orthogonal system for which the degree, D_c, is zero. In this case, $\lambda_i = 1$ for all i = 1, 2, ..., p, and we have

$$\frac{d}{dk}\,\mathrm{MSE} = -2\Sigma\frac{\lambda_i}{(\lambda_i + k)^3}\,[\sigma^2 - k\alpha^2] = 0$$

$$\Sigma(\sigma^2 - k\alpha_i^2) = 0$$

$$p\,\sigma^2 - k\Sigma\alpha_i^2 = 0$$

$$\therefore k = k_o = \frac{p\sigma^2}{\Sigma\alpha_i^2} = \frac{p\sigma^2}{\Sigma\beta_i}$$

This strange phenomenon has one important advantage and one important disadvantage. The advantage is that if we accept the minimum MSE criterion as a measure of the goodness of an estimator, for any problem, the OLS procedure is always inferior. The disadvantage is that it renders Hoerl and Kennard's ridge procedure based on minimum MSE rather absurd.

2.2.16 For any problem the optimal k depends on the true regression coefficient vector, β, and the variance of the residual of the linear model, σ^2.

Proof: From Eq. (2.39), at minimum MSE

$$\frac{d}{dk}\,\mathrm{MSE} = -2\Sigma\frac{\lambda_i}{(\lambda_i + k)^3}\,(\sigma^2 - k\alpha_i^2) = 0 \tag{2.43}$$

Although the explicit form of the optimal k cannot be solved from this complex non-linear equation, it is obvious that the optimal k is a function of the true regression coefficient vector, α or β, and the variance of the residual of the linear regression model; that is, σ^2.

Comment: The multicollinearity problem arises from the interdependency among the explanatory variables, not from the dependency between the dependent variable y, and the explanatory variables X's. Therefore, if the task of ridge regression is to reduce the harmful effects of multicollinearity, the optimal k should not depend on any parameter which depends on the y variable such as β or σ^2; that is, the optimal k should be a nonstochastic parameter.

2.3 The Optimal k

It has been proven (property 15 of ridge regression) that the optimal k is the one which gives the MSE for the data on hand. It has also been proven (property 14) that for any problem there is one optimum k, and a wide range of k, $0 < k < k_{max}$, which give smaller MSE than that of OLS regression. Unfortunately the optimal k depends on the true regression coefficient vector β, and the variance of the residual σ^2, in the linear regression model. These two parameters, however, are population parameters not universal constants. Thus, because of this dependency, the optimal k cannot be calculated. Instead it has to be estimated from the sample data. So far more than fifteen methods have been described; for example, Allen (1974), Guilkey and Murphy (1975), Hemmerle (1975), Hemmerle and Brantle (1978), Hocking et al. (1976), Hoerl and Kennard (1970a, 1976), Hoerl, Kennard, and Baldwin (1975), Kasarda and Shih (1977), Lawless and Wang (1976), McDonald and Galarneau (1975), Obenchain (1975), and Vinod (1976). Each of these methods has its own advantages and disadvantages. Unfortunately, however, no one method can guarantee to give a better k than any other; and none can even guarantee a smaller MSE compared to that of OLS regression. These difficulties have marred the superiority of ridge regression over OLS regression.

Nevertheless, in the next chapter three of the more distinctive methods of estimating the elusive optimal k will be discussed; namely, Hoerl and Kennard's ridge trace method, Kasarda and Shih's method, and Vinod's Index of Stability method. The advantages and disadvantages of each will be discussed.

CHAPTER THREE

3.0 SIMPLE RIDGE REGRESSION IN PRACTICE

3.1 Hoerl and Kennard's Ridge Trace Method

In the earliest version of simple ridge regression, Hoerl and Kennard (1970a) determined the "optimal k" visually from the "ridge trace"; that is, from the plot of $\hat{\beta}_i^*$'s and the residual sum of squares as functions of k. Figure 3.1 below shows the ridge trace of a 10 factor problem obtained from Hoerl and Kennard (1970b).

Figure 3.1 About Here

The ridge trace depicts the effects of multicollinearity on each of the regression coefficients. From the trace it can be seen that when k increases the effect of multicollinearity is dampened and the regression coefficients are stabilized. The optimal k is then selected visually at the region which starts to give stabilized regression coefficients.

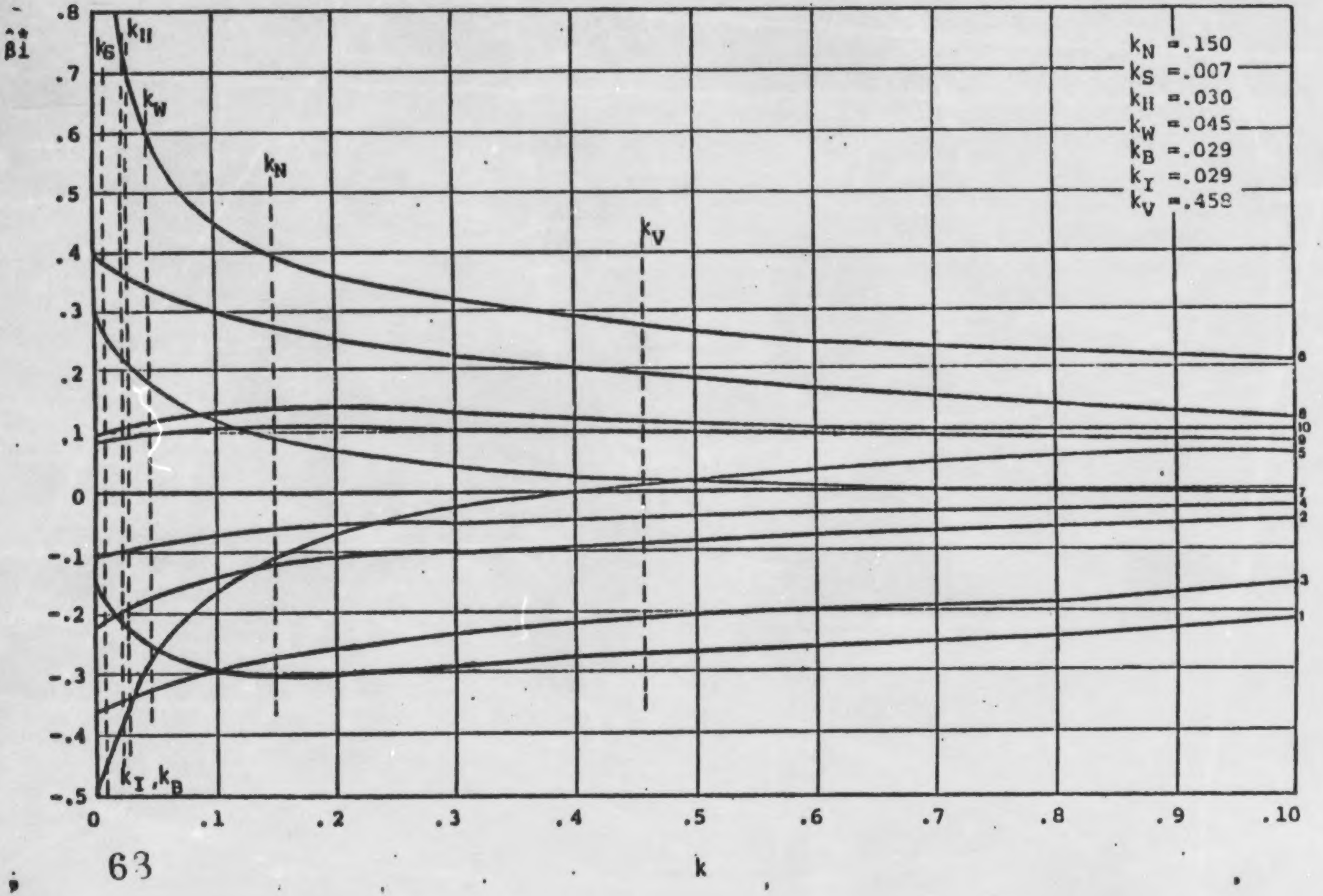
Four guidelines are given by Hoerl and Kennard (1970a) in order to aid the analyst in the selection of the optimal k.

1. At a certain value of k the system will stabilize and at that point will demonstrate the general characteristics of an orthogonal system.

2. Coefficients will not have unreasonable absolute values with respect to the factors for which they represent rates of change.

Figure 3.1 10-factor example from Hoerl & Kennard (1970 b)

(N=36, V_{max}=9.11, D_{max}=.952)



3. Coefficients with apparently incorrect signs at $k = 0$ will have been changed to the proper sign.

4. The residual sum of squares will not have been inflated to an unreasonable value. In particular it will not be large relative to the minimum residual sum of squares, or large relative to what would be a reasonable variance for the process generating the data.

It is obvious that these four guidelines are vague, subjective, need prior knowledge of the regression coefficients and would prove difficult to apply in an unambiguous fashion. Furthermore, the trace appears to be more stable at higher k; hence has a tendency to lead the analyst to select a k that might be too high.

Because of these drawbacks, the obtained "optimal k" cannot guarantee to give estimates that are better than the OLS ones. Nevertheless, in spite of these limitations, the ridge trace is still a useful plot. It distinctively displays the characteristics of the explanatory data set and the effectiveness of the ridge procedure in stabilizing the regression coefficients. It can also be used to check the optimal k estimated by using other methods to see whether they fall in the stable region of the ridge trace as desired by a good estimator.

3.2 Kasarda and Shih's Method

It has been shown (proposition 15, chapter 2) that theoretically the true optimum k is one which minimizes the total mean square error of the estimates of regression coefficients. The MSE, however, depends on the unknown true regression coefficients (see property 8) and the variance of the residuals (see property 16). Mathematically or technically it is not difficult to obtain the optimal k from Eq. (2.43), even for very high p, if the two parameters are known.

Thus, Kasarda and Shih (1977) have argued that since the OLS estimates of σ^2 and β, under the normality assumption are unbiased and consistent; and, further, since $\hat{\beta}$ has the minimum variance among all unbiased estimators, then the two OLS estimates, σ^2 and $\hat{\beta}$, may be used to replace their true values in order to obtain the optimal k by minimizing the OLS estimate of MSE; which is written as

$$\hat{MSE} = \Sigma \hat{Var}\,(\hat{\beta}_i^*) + \hat{Bias}^2(\hat{\beta}^*)$$

$$= \hat{\sigma}^2\ \Sigma\ \frac{\lambda_i}{(\lambda_i + k)^2} + k^2\ \frac{\hat{\alpha}_i^2}{(\lambda_i + k)^2}$$

The validity of this method obviously rests on the validity of replacing σ^2 and α by their OLS estimates. The replacement of σ^2 is not problematic as explained by Johnston (1972, p. 163), but the replacement of α (or β) by $\hat{\alpha}$ (or $\hat{\beta}$) definitely is. This is because the OLS estimates of α (or β) could be far off due to the high degree of multicollinearity in the problem as explained under harmful effect (2) in chapter 1. Thus, in the presence of a high degree of multicollinearity in the problem the replacement of α by $\hat{\alpha}$ would definitely produce too large a square bias in the estimate of MSE, and this in turn would produce an estimate of k (called k_s) smaller than the true optimal k (k_o), due to the fact that the variance component in MSE is a monotonic decreasing function and the square bias component is a monotonic increasing function with $\beta'\beta$ as its upper limit. In the absence of rigorous proof, the above argument can be depicted graphically. See Figure 3.2 below.

Figure 3.2 About Here

In summary, the Kasarda and Shih (1977) method is an effort to overcome the subjective features of the ridge trace method. Yet, because the MSE is dependent on the true unknown parameters σ^2 and α it can only be obtained by using inaccurate OLS estimates of σ^2 and α. Thus, the estimating procedure has two disadvantages. First, it necessarily produces a k which is an underestimate of the true optimal k; and the higher the degree of multicollinearity, the worse the estimation. Second, it produces a stochastic k, while optimal k should be nonstochastic due to the fact that multicollinearity is a nonstochastic problem as explained in the discussion centering on property (16) of ridge regression.

There is a sense, however, in which the first disadvantage might be regarded as an advantage. This is because

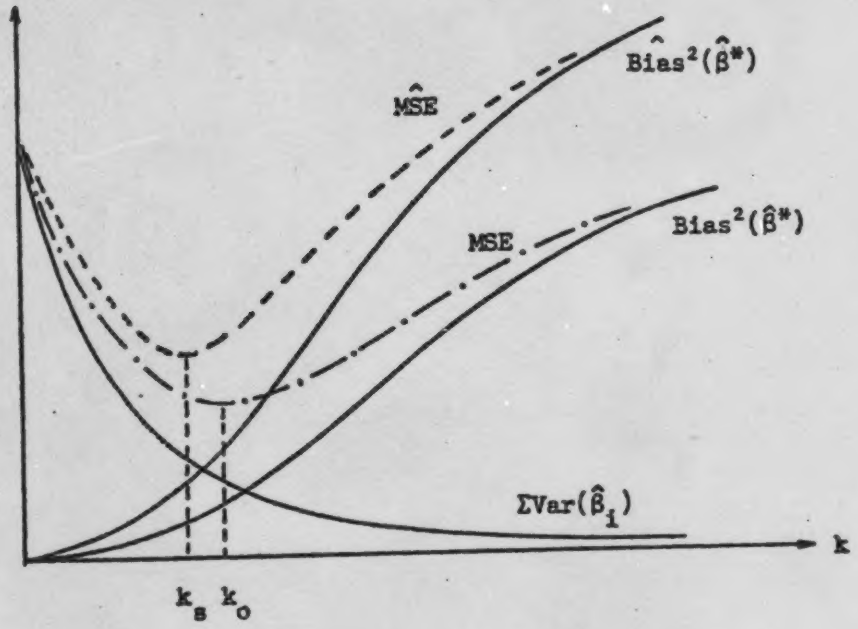


Figure 3.2 The mean square error function MSE and its OLS estimate, $\widehat{\text{MSE}}$, where k_s is Kasarda & Shih's k, and k_o is the true optimum k.

k_s will never be an overestimate of k_o. Thus, unlike some of the other ridge regression methods for estimating k which produce too high a bias the Kasarda and Shih method will always produce an estimate with smaller MSE than that of OLS.

3.3 Vinod's Index of Stability Method

As pointed out above (chapter 2, property 4) the ridge estimator $\hat{\beta}^* = (X'X + kI)^{-1}X'y$ has a variance-covariance matrix of

$$\mathrm{cov}(\hat{\beta}^*) = \sigma^2[\mathrm{VIF}]$$

where

$$\mathrm{VIF} = P\ \mathrm{Diag}\left[\frac{\lambda_i}{(\lambda_i + k)^2}\right] P'$$

is the variance inflation factor matrix. For a completely orthogonal system, $\lambda_1 = \lambda_2 = \ldots = \lambda_p = 1$, it can be easily seen that the VIF matrix is equal to a constant matrix with elements

$$\frac{\lambda_i}{(\lambda_i + k)^2} \qquad \left(\text{i.e., } \frac{1}{(1 + k)^2}\right)$$

and, therefore, the matrix

$$\frac{p}{\Sigma \frac{\lambda_i}{(\lambda_i + k)^2}} \quad P\ \mathrm{Diag}\left[\frac{\lambda_i}{(\lambda_i + k)^2}\right] P'$$

is an identity matrix.

For a nonorthogonal system the above property of the VIF will not be satisfied, and the absolute values of the elements will be large. This suggests a numerical measure, which Vinod (1976) called the Index of Stability of Relative

Magnitude (ISRM) of $\hat{\beta}^*$, defined by

$$\text{ISRM} = \Sigma[{}_p\lambda_i/(\lambda_i + k)^2 S - 1]^2$$

where

$$S = \Sigma\lambda_i/(\lambda_i + k)^2$$

This numerical measure represents the quantification of Hoerl and Kennard's (1970a) concept of stability which will be zero for a completely orthogonal system. Since in ridge regression it is desired to minimize the effects of the nonorthogonality of the system, the index of stability (ISRM) should be minimized; that is, for optimal k, ISRM has a minimum value.

Because of the complexity of the stability function it has proven impossible to solve for the optimum k as an explicit function of λ_i, and therefore it has either to be solved graphically by plotting the ISRM as a function of k, or by using an iterative approach.

Vinod's method at first sight may seem to be one of the best compared to most of the ridge regression methods proposed. However, as will be shown below in chapters 5, 7, and 8 its application has not proved very satisfactory. Nevertheless, it has three important advantages: (i) it quantifies Hoerl and Kennard's concept of stable region and to that extent estimates k objectively; (ii) it gives a more definitive k than that generated by the ridge trace method; and (iii) it gives a nonstochastic k.

There are three disadvantages which seem in terms of application to outweigh the advantages. First, it was not proven that the optimal k gives the minimum mean square error as required by the criterion of ridge regression. Second, there is no guarantee that the optimal k obtained will not be larger than k_{max} (see property 16 in chapter 2), as required by any valid ridge procedure. Third, in some cases, the index of stability gives more than one minimum point, while in theory there should only be one optimal k that gives the minimum MSE as shown by property 15 (chapter 2). To illustrate this problem the multiple minimum of ISRM for the six factor example given by Vinod (1976) has been graphed in Figure 3.3.

Figure 3.3 About Here

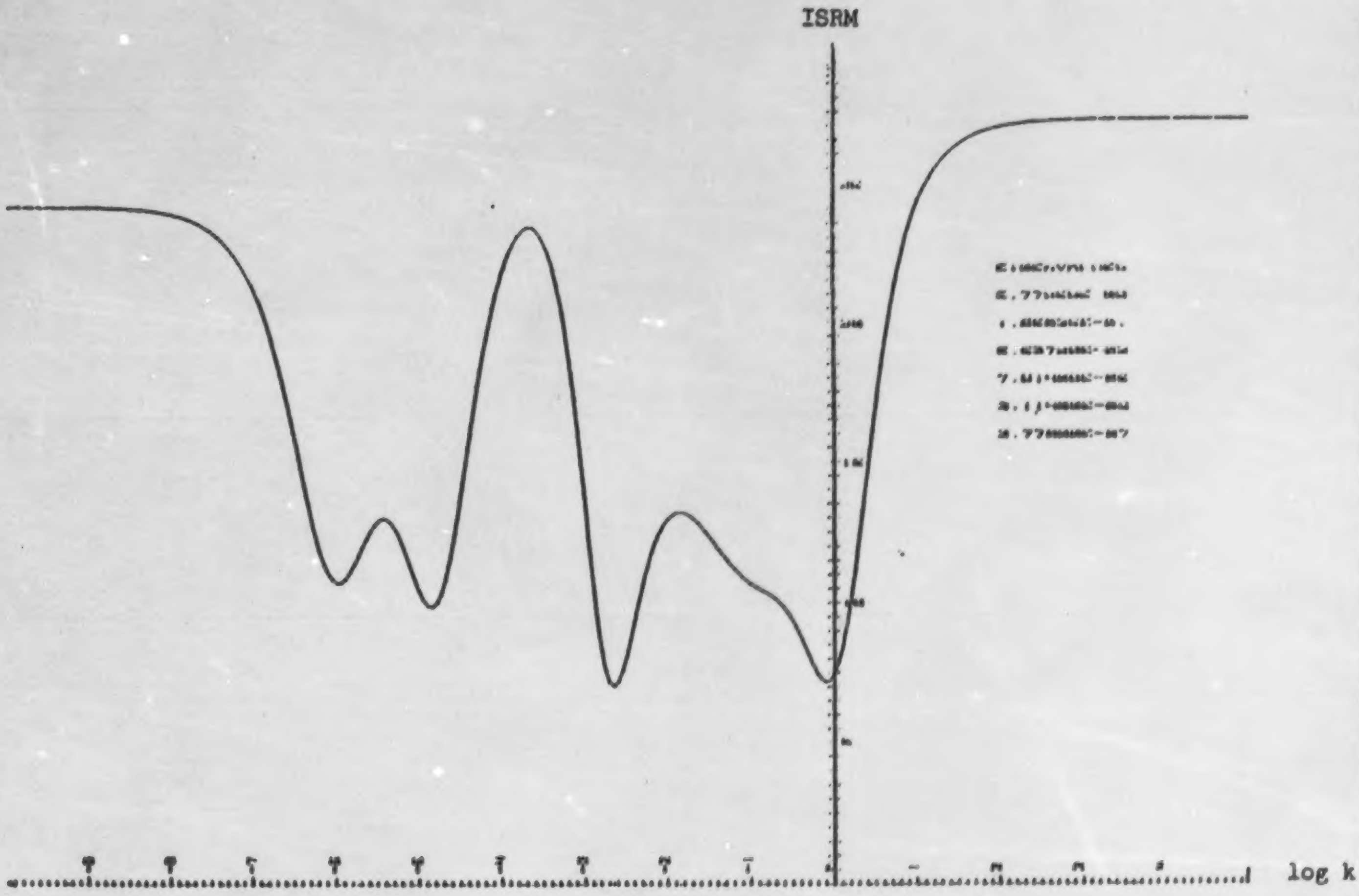
3.4 Advantages and Disadvantages of Simple Ridge Regression

The prime purpose of this monograph is to develop an estimating procedure of the biased, reduced-variance variety designed to cope satisfactorily with the six harmful effects of multicollinearity described in the first chapter (section 1.2). Ridge regression is one class of biased, reduced-variance estimators which go most of the distance towards overcoming these harmful effects as was shown through the discussion of properties of ridge regression in chapter two.

First, although OLS estimates have large sampling variance the ridge procedure produces smaller sampling variance. The proof of this assertion is given under property (4) of ridge regression in chapter two. Secondly, it is

Figure 3.3

The multiple minimum of ISRM for the 6-factor example from Vinod (1976)



because OLS estimates have large sampling variance that they are sensitive to sampling fluctuation. And because ridge regression produces smaller sampling variance the ridge estimate is consequently less sensitive to sampling fluctuation.

Least squares estimates are sensitive to model misspecification; that is, to stochastic error. But since stochastic error is a source of the total variance of the estimated regression coefficients (see the discussion of random error in section 1.4.1 above), then the ridge solution is a third source of advantage because ridge regression reduces the variance of the regression coefficients (see property 4) which, therefore, renders the ridge solution less sensitive to model misspecification than OLS solutions. Fourth, since the higher the variance of the estimates the higher the dependence of the regression coefficients -- that is, the more likely the estimates are to be correlated; hence, erroneous -- any procedure reducing the variance such as ridge regression will generate more accurate estimates.

Fifth, in ill-conditioned data sets OLS coefficient vectors are often too long; hence, "off" in both magnitude and direction. But it was proven in the discussion of property one (2.2.1) of ridge regression that the ridge method generates a shorter coefficient vector than OLS. Sixth, and finally, when a data set is ill-conditioned, OLS models do not provide accurate prediction; not even when the coefficient of determination (R^2) is high. But even though ridge regression produces smaller multiple R^2 than OLS regression (for proof see property 11) the ridge procedure always produces a more accurate prediction equation if the bias is not too large. For proof see property (13) of ridge regression.

Though these are compelling reasons for using ridge regression in preference to OLS regression they are rendered largely academic when ridge regression is evaluated in the light of its disadvantages. Five of the properties of ridge regression are problematic; namely, properties six through eight, and properties fifteen and sixteen. Here, we shall be concerned with the last three. From property eight it is noted that the total MSE of ridge regression depends on the true and unknown parameters σ^2 and α, which means effectively that the total MSE cannot be obtained as was noted in the discussion under the Kasarda and Shih method. From property fifteen it is noted that for any regression problem there exists a positive "optimal k" (k_o) which is (absurdly) true even for an orthogonal system for which the multicollinearity is zero. Third, from property sixteen it is obvious that the optimal k depends on the true regression coefficient vector, β, and the variance of the residual of the linear model, σ^2 which are in turn dependent on the y-variable. Thus, k depends on the interdependencies between the dependent and independent variables -- that is, on the dependency between the y-variable and the x-variables. Yet, multicollinearity arises solely from the interdependency among the explanatory variables, X's. This means that in ridge regression estimation k is a stochastic parameter when in fact (and theory) it should be a nonstochastic one.

The four dilemmas of ridge regression stem from these three properties; namely, (i) the theoretical value of optimal k is stochastic when it should be nonstochastic if the task is to reduce the harmful effects of multicollinearity; (ii) the true optimal k depends on unknown population parameters, a fact which renders the problem unsolvable; (iii) even for an orthogonal system there is an optimum k at which MSE is minimum; and (iv) the bias of the ridge estimates can never be obtained. Effectively the optimum k cannot be calculated; consequently, the performance of ridge regression cannot be evaluated.

The question arises as to whether an estimating procedure can be devised which retains the advantages of the ridge regression procedure while overcoming its inherent disadvantages. The authors think there is. Their alternative, called normalization ridge regression analysis is the subject of the next chapter.

CHAPTER FOUR

4.0 SIMPLE NORMALIZATION RIDGE REGRESSION ANALYSIS

4.1 Introduction

Four dilemmas of ridge regression were identified in section 3.4 above. A careful examination of properties eight, fifteen, and sixteen of ridge regression (section 2.2) indicates that all four dilemmas have a common source; namely the estimation of the minimum mean square error. They are attributable to the fact that because the MSE is a function of population parameters β and σ^2, it is a stochastic function which cannot be accurately estimated.

In this chapter we use a weaker criterion than the MSE criterion to estimate the biasing parameter k; that is, weaker in the sense that it is limited to fewer tasks than that of MSE. It will be shown that through the use of this criterion, called the variance normalization criterion, all but the last of the four dilemmas - namely, obtaining the bias of the ridge estimates - will be avoided, while at the same time retaining all the previously noted advantages of ridge regression which were summarized in section 3.4.

In Chapter One (section 1.4) the standard solutions to the multicollinearity problem were discussed. Here it was shown that the total variance of the estimated regression coefficients based on the OLS procedure could be expressed as:

$$\Sigma \mathrm{Var}(\hat{\beta}) = \sigma^2 \mathrm{Tr}(X'X)^{-1}$$

$$= \frac{\varepsilon'\varepsilon}{n-p}\mathrm{Tr}(X'X)^{-1} \qquad \text{(See Eq. 1.12)}$$

The total variance depends, therefore, on three independent factors - the random error ($\varepsilon'\varepsilon$), the sample size or the degree of freedom of the sum of squares error (n-p), and the degree of multicollinearity, $\mathrm{Tr}(X'X)^{-1}$. Thus, for any multiple regression problem the total variance of the estimated coefficients may be reduced by improving the measurement, by improving model specification, by increasing the sample size, and by reducing the inflation of variance due to multicollinearity.

Though the ridge regression method developed by Hoerl and Kennard (1970a) was originally intended to do just one task; namely, to reduce the inflation in the variance of the estimates due to multicollinearity, use of the MSE criterion for determining k is effectively to use an omnibus procedure. This is because ridge regression in suppressing the VIF to less than unity is also suppressing the variance attributable to other causes of variance; namely, measurement error, model incompleteness or system misspecification and small sample size. It is this omnibus or multifunctional aspect of the procedure which has forced ridge regression to be stochastic (that is, dependent on β and σ^2) when it should be non-stochastic, and why, even for an orthogonal model, there is an optimum k which generates minimum MSE. See the comment under property 15 in Chapter Two.

4.2 The Normalization Criterion

From the preceding argument it is obvious that if we want to limit the simple ridge regression procedure to the performance of a single task; namely, that of reducing the inflation of variance due to multicollinearity, the lowest permissible total variance is $p\sigma^2$. This is the value for the best condition; namely, that present in an orthogonal system. That is,

$$\Sigma \mathrm{Var}(\hat{\beta}^*) = p\sigma^2$$

but
$$\sigma^2 \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} = p\sigma^2$$

and, therefore

$$\frac{1}{p} \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} = 1 \qquad (4.1)$$

In the common language we can say that in order to perform the single task of reducing variance inflation due to multicollinearity, and only multicollinearity, one should normalize the variance inflation factor (VIF) so that it is equal to one. The value of k (called k_N) which satisfies this condition is the k which satisfies Eq. (4.1).

If by using this procedure, the resultant variance is still too large for the practical application of the regression model; and if it is desirable to further reduce the variance, it should be accomplished by improving the measurement, improving the model specification, and increasing the sample size; and not by further suppression of the variance inflation factor. Due to the seriousness of multicollinearity the analyst should locate the sources of the problem (Mason, Webster and Gunst, 1975) in order to try to eliminate them if physically possible. In the event that this is not possible and in the event that a principal component "solution" (Maddula, 1977, pp. 193-194) is undesirable, a biased, reduced-variance method such as ridge regression or variance normalization may be used instead in an effort to reduce (but not eliminate) the effects of the multicollinearity problem without inducing too much bias. It is far better, however, to eliminate the causes than to use a statistical solution which may be interpreted as merely the treatment of the symptom.

4.3 The Normalization Criterion versus the Mean Square Error Criterion

Though the important advantages of ridge regression were identified above (section 3.4) it was found that the procedure was flawed in four ways: first, the true optimum k (k_o) could not be calculated; second, k_o is stochastic when it should be nonstochastic; third, the procedure does not satisfy the boundary condition - that is, even for orthogonal data ridge regression estimates are different than OLS estimates; and fourth, the procedure is multifunctional.

Ridge regression is multifunctional because it accomplishes four tasks. In suppressing the variance inflation factor to less than unity it is also suppressing the variance due to measurement error, model misspecification and small sample size, as well as reducing variance inflation caused by multicollinearity. As explained in section 2.2 (property sixteen) it is this multifunctional aspect of the procedure which forces the k parameter to be stochastic. How well the ridge regression method accomplishes these tasks depends entirely on the size of the estimated k. In both ridge regression and normalization regression methods one assumption is that k will be less than k_{max}, which is the k-value for which the reduction in the total variance is equal to the total squared bias introduced by the ridge procedure. In terms of the effectiveness index, Eft, k_{max} is the value which gives Eft = 1 (see property 14 in Chapter 2). As is pointed out in 4.4.6 below, however, the assumption may not be correct.

Nevertheless, in spite of this limitation the normalization criterion has several advantages. First, the parameter k_N can be calculated accurately. It is non-stochastic, as required. In addition the average variance inflation factor will never be suppressed. In Chapter Five (section 5.3) it will be shown that normalization regression is more conservative in terms of several performance indices than several ridge regression methods. Finally, normalization regression estimation may help the analyst narrow down or even locate the main source or sources of variance inflation in a model.

This last claim requires substantiation. Suppose a model is estimated using a large data set; and suppose, further, that from other sources we know the model is well specified. If after analysis using normalization regression the variance is still too large, it can be inferred that the major source of this variance is measurement error. This follows from the fact that normalization regression performs the single task of reducing the degree of multicollinearity.

Consider a second example. Suppose an analyst has a data matrix corrected for attenuation; that is, one which it is assumed is composed of perfectly reliable variables. Suppose, further, that the data set had a large number of cases. If after analysis using normalization regression the variance inflation is still too large, then the major source of inflation is probably stochastic error attributable to model misspecification.

In the first example the analyst may cope satisfactorily with the problem by improving the measures of the variables, and in the second example the problem might be resolved by improving the design of the model. It is conceivable that normalization regression analysis could be used not merely to statistically reduce the degree of multicollinearity in a model, but also to identify the specific sources of variance inflation. In this way an analyst could decide which strategy to adopt to cope more satisfactorily with the general problem of variance inflation of the model parameters.

4.4 Performance Indices

How can the performances of the various ridge regression methods be compared? Different performance indices have been used by different analysts. Hoerl and Kennard (1970a) and Marquardt and Snee (1975), for example, compare the different biasing parameters, the "k's", the inflated residual sums of squares, ϕ's, and the maximum variance inflation factors, $V_{max's}$. McCabe (1978) and Obenchain (1977) have used the α-acceptability index. In section 1.3 above use of the D_{max}-value was proposed. Four additional performance measures will also be used; namely, a sensitivity ratio, the relative reduction in variance, the average percentage error of estimated variance, and the traditional standby - the coefficient of determination. In sum, nine performance indices may be used to evaluate the effectiveness of ridge regression procedures.

4.4.1 The biasing parameter, k. Simple ridge regression involves augmenting the diagonal of the normal equation matrix with a small positive quantity, called "k" (a constant), in order to produce regression coefficients with smaller variance at the expense of introducing a small bias. As shown in Chapter Three, however, the main difficulty with ridge regression is the estimation of the optimal k (k_o) which produces the minimum total mean squared errors. Though in theory there always exists an optimal k for any regression problem, in practice the optimal k depends on two unknown population parameters, the error variance, σ^2, and the true regression coefficient vector, β. Thus, an optimum value of the magnitude of k cannot be obtained though numerous approximations of the k_o-value are possible. It is for this reason - the dependence of the bias on population coefficients - that it is not possible to use a single measure such as MSE, Eft or any other measure as a performance index. Rather, comparisons have to be made between various performance indices, including "k".

Note the difference between the family of k estimates using the minimum MSE criterion on the one hand and the k estimate using the variance normalization criterion on the other. With the minimum MSE criterion, the optimum k's are those which give the smallest MSE of the estimates; while with the variance normalization criterion, the k's are those that set the variance of the individual estimates $\hat{\beta}_i^*$ equal to their theoretical value, the error variance σ^2. The variance normalization k, then, normalizes the variance inflation factors contributed by multicollinearity to unities.

4.4.2 Inflation in the residual sum of squares, ϕ. OLS regression is based on the minimum residual sum of squares criterion. As noted in Chapter Two (2.1) the use of a different criterion such as MSE produces an inflated residual sum of squares.

4.4.3 The maximum variance inflation factor, V_{max}. The maximum variance inflation factor is a linear measure of the severity of multicollinearity in the data set. For a perfectly orthogonal data set it is equal to one; and for perfect collinearity it is infinity.

4.4.4 The acceptance level of a ridge estimate, α. This is defined by McCabe (1978) as the significance level of the F-ratio.

$$F_{p,n-p,(1-\alpha)} = \frac{(\hat{\beta}'\hat{\beta}^*)'X'X(\hat{\beta} - \hat{\beta}^*)}{p\sigma^2} \qquad (4.2)$$

A ridge estimate is called "α-acceptable" if the estimate is in the (1 - α) 100% confidence region; that is, if the estimate is not sufficiently different from the OLS estimate. We do not agree, however, with McCabe's notion that the best ridge estimates ought not to be significantly different from the OLS estimates.

4.4.5 The degree of multicollinearity, D_{max}. It was shown in Chapter One (1.3) that the D_{max} measure of multicollinearity has a range of zero to one; $D_{max} = 0$ for no multicollinearity, and $D_{max} = 1$ for perfect multicollinearity.

4.4.6 The empirical sensitivity ratio, E.S. Ratio. The empirical sensitivity ratio is the ratio of the sum of the squares of the fluctuation of estimated ridge regression coefficients (SSF*) to those of OLS regression (SSF); that is,

$$\text{E.S. Ratio} = \frac{SSF^*}{SSF}$$

$$= \frac{\Sigma(\hat{\beta}_i^* - \tilde{\beta}_i^*)^2}{\Sigma(\hat{\beta}_i - \tilde{\beta}_i)^2} \qquad (4.3)$$

where $\tilde{\beta}_i$ and $\tilde{\beta}^*_i$ represent the ith perturbed OLS regression coefficient and that of the ridge regression coefficient respectively; and where the perturbed coefficients are obtained by introducing a small perturbation Δr to the largest correlation coefficient in the correlation matrix X'X. The amount of the perturbation used in the analysis is $|\Delta r| = 0.01$, which is a measurable approximation to the precision of a correlation coefficient in most measurement.

4.4.7 Relative reduction in variance (RVAR). RVAR is a measure of the percentage reduction of the total variance in the parameter estimates generated by the ridge procedure over the OLS procedure. By definition it can be expressed as:

$$\text{RVAR} = \frac{\sigma^2 \text{tr}(X'X)^{-1} - \sigma^2 \text{tr}[\text{VIF}]}{\sigma^2 \text{tr}(X'X)^{-1}}$$

$$= 1 - \frac{\text{tr}[\text{VIF}]}{\text{tr}(X'X)^{-1}} \qquad (4.4)$$

where $[\text{VIF}] = (X'X + kI)^{-1}\, X'X(X'X + kI)^{-1}$

4.4.8 The average percentage error of estimated variance (Av.%Er.VAR). The average percentage error of the estimated variance of the $\hat{\beta}^*_i$'s is defined as:

$$\text{Av.\%Er.VAR} = \frac{1}{p} \sum_{i=1}^{p} \left| \frac{\text{Var}(\hat{\beta}^*_i) - \sigma^2}{\sigma^2} \right|$$

where $\text{Var}(\hat{\beta}^*_i)$ is the estimated variance of $\hat{\beta}^*_i$ and σ^2 is the error variance of the regression model. Since $\text{Var}(\hat{\beta}^*_i) = \sigma^2[\text{VIF}]_{ii} = \sigma^2\text{VIF}_i$, it is obvious that

$$\text{Av.\%Er.VAR} = \frac{1}{p} \Sigma |\text{VIF}_i - 1| \qquad (4.5)$$

where VIF_i is the variance inflation factor of $\hat{\beta}_i^*$. This performance index is a measure of the average inflated variance that has not been "eliminated" (including suppressed variance) by the respective ridge procedure. It can be used to compare the performances of various ridge procedures and can also be used to compare the performance of the same ridge procedure on different data sets. The lower the average percentage error of estimated variance the more the inflated variance due to multicollinearity will have been removed. If, for example, the Av.%Er.VAR = 0, the inflated variances of the estimates due to multicollinearity will have been completely removed.

4.4.9 The coefficient of determination (R^2). The square of the multiple correlation coefficient or the coefficient of determination, is equal to the proportion of the Y variance explained by the linear influence of the predictor variables, X's. A multiple R value of 0.9, therefore, indicates that the least squares regression of Y on the X's accounts for 81 per cent of the variance in Y. Analysts usually associate a high R^2 value with a good fit and a low value with a poor fit, but in the presence of multicollinearity this is not the case. Thus, we have serious reservations about the conventional interpretation (see 2.2.11 and 2.2.13 above).

CHAPTER FIVE

5.0 RIDGE REGRESSION AND NORMALIZATION RIDGE REGRESSION: SOME COMPARISONS

5.1 Introduction

In chapters one through three we have defined the multicollinearity problem; discussed several standard solutions; and noted that simple ridge regression in a variety of forms has become the most influential statistical solution. The properties of simple ridge regression were also defined and three of the standard ridge regression techniques examined in detail. An assessment of some of the theoretical advantages and disadvantages of ridge regression analysis was also made.

In chapter four the theory underlying a new ridge regression procedure called normalization ridge regression was developed. It was shown that, in theory, normalization ridge regression analysis corrected three of the four dilemmas of simple ridge regression without adversely affecting its several important advantages. What has not been done is to demonstrate the performance qualities of the new method. This is the purpose of the present chapter. The task calls for the making of comparisons -- comparisons between OLS regression, selected ridge regression methods, and normalization ridge regression in terms of their respective performance qualities.

Chapter five, then, completes our discussion of multicollinearity by examining the authenticity of normalization ridge regression through the comparative analysis of four nonorthogonal data sets. The question asked is: Do normalization ridge regression estimates in practice compare favourably or not with the other methods? Four models selected from the ridge regression literature will be estimated. Eight different estimating methods will be used -- six ridge regression methods, the normalization ridge regression, and ordinary least square regression. Thirty-two analyses in all.

The chapter has four parts. First, the four models from the ridge regression literature are briefly described. Second, the models are estimated using eight different methods and the parameter estimates from the different methods are compared for each model. Third, the performance characteristics of each method are compared as are the ridge traces for each set of results. Fourth, the overall patterning of the results of these comparisons is described.

5.2 The Models

5.2.1 *The ten-factor example from Hoerl & Kennard (1970b).*

Though the origins of ridge regression analysis are usually traced to the papers by Hoerl and Kennard (1970a, 1970b) that appeared in Technometrics, three earlier papers by Hoerl (1959, 1962 and 1964) in Chemical Engineering Progress referred to "ridge analysis" and were the seminal sources (cf. Hocking, 1976, p. 47). Here, we are interested in the ridge regression analysis of a ten-factor model which was used by Hoerl and Kennard (1970b) to illustrate the practical use of ridge regression.

The source of Hoerl and Kennard's example was the Gorman and Toman (1966) paper describing the strategies for selecting the subset of factors for obtaining the "best" regression equation. The problem was described as follows: "The ten-variable example concerns the search for an interpolation formula to fit data taken from routine operation records for a petroleum refining unit." (Gorman & Toman, 1966, p. 43)

The data input for the problem is presented below in X'X and X'y correlational form. Several large interfactor correlations are reflected in the X'X eigenvalues; also shown in Table 5.1. Since

$$E(L_i^2) = \sigma^2 \Sigma \lambda_i^{-1} > \sigma^2 \lambda_p^{-1} \qquad (5.1)$$

where $E(L_i^2)$ is the expected squared distance of the least squares coefficient estimate $\hat{\beta}$ from β; and since $\Sigma\lambda_i^{-1} = 33.825\sigma^2$, then the distance is $\Sigma\lambda_i^{-1}/p = 33.825/p = 3.4$ times greater than what it would be for an orthogonal system. (Hoerl & Kennard, 1970b, p. 70)

Table 5.1 About Here

5.2.2 The fifteen-factor air pollution problem from McDonald and Schwing (1973). The problem is as follows:

> Recently, Lave & Seskin (1970) and Hickey et al. (1970) using multiple regression analyses on large data banks of annual mortality rates and pollution measurements have provided a link between long-term, high pollution levels (sulphates, particulates, and heavy metals emanating primarily from stationary sources) and increased mortality....To eliminate the fact that pollution levels can be correlated with other variables which can influence health, it is often necessary to investigate several highly correlated (non-orthogonal) variables simultaneously. (McDonald & Schwing, 1973, p. 463)

The problem, then, involves using ridge regression in an attempt to obtain more accurate predictor variable effects for nonorthogonal data sets than is possible using standard OLS regression methods. The variables for the air pollution

Table 5.1

Correlation Matrix and Eigenvalues of X'X and X'y for the Ten-Factor Example from Hoerl & Kennard (1970b) (N = 36)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1.00									
X_2	-0.04	1.00								
X_3	.51	-0.0	1.00							
X_4	.12	-0.16	0.0	1.00						
X_5	-0.71	.06	-0.59	-0.07	1.00					
X_6	-0.87	.09	-0.65	-0.09	.84	1.00				
X_7	-0.09	.24	-0.02	.03	.38	.13	1.00			
X_8	-0.0	.01	.34	.08	-0.36	-0.20	-0.48	1.00		
X_9	-0.04	.09	-0.08	.02	-0.14	.04	.07	-0.18	1.00	
X_{10}	-0.36	-0.30	-0.44	-0.09	.54	.45	.40	-0.46	.05	1.00
Y	-0.81	-0.10	-0.63	-0.10	.56	.81	.04	.06	.16	.45
λ_i^a	3.692	1.542	1.293	1.046	0.972	0.659	0.357	0.220	0.152	0.068

a. $\Sigma\lambda_i^{-1} = 33.825$

problem are described in Table 5.2, and the summary input data in X'X and X'y form are reproduced in Table 5.3 Several large intercorrelations are reflected in the small eigenvalues of X'y also shown in Table 5.3. Application of Eq. (5.1) shows that the expected squared distance of the least squares coefficient estimate $\hat{\beta}$ from β is $263.06\sigma^2$ which is more than seventeen times $(263.06\sigma^2/15\sigma^2)$ what it would be for an orthogonal system.

Tables 5.2 and 5.3 About Here

5.2.3 *The ten-factor gas mileage study from Hocking (1976)*. Hocking's ten-factor model was formulated in order to predict gasoline mileage for 1973-74 automobiles. The predictors described in Table 5.4 were selected physical characteristics of various types of automobiles. The correlation matrix and eigenvalues of X'X are presented in Table 5.5. Application of Eq. (5.1) to this input data shows that the expected squared distance of the least squares coefficient estimate $\hat{\beta}$ from β is $95.25\sigma^2$, which is more than nine times what it would be for an orthogonal model.

Tables 5.4 and 5.5 About Here

5.2.4 *The five-factor employee satisfaction example from Price (1977)*.

Price's problem concerns the identification of the factors determining employee satisfaction with their supervisors. The dependent variable Y was the employee's assessment

Table 5.2

Description of Variables in the Fifteen-Factor Air Pollution Problem from McDonald & Schwing (1973)

Variable	Description
X_1	Mean annual precipitation in inches
X_2	Mean January temperature in degrees Fahrenheit
X_3	Mean July temperature in degrees Fahrenheit
X_4	Percent of 1960 SMSA population aged 65 or over
X_5	Population per household, 1960 SMSA
X_6	Median school years completed for persons over 25 in 1960 SMSA
X_7	Percent housing units with all facilities
X_8	Population per square mile in urbanized area in 1960
X_9	Percent of 1960 urbanized population non-white
X_{10}	Percent employment in white collar occupations in 1960
X_{11}	Percent of families with income LT $3000 in 1960
X_{12}	Relative pollution potential of hydrocarbons, HC
X_{13}	Relative pollution potential of oxides of nitrogen, NO_x
X_{14}	Relative pollution potential of sulfur dioxide, SO_2
X_{15}	Percent relative humidity, annual average at 1 p.m.
Y	Total age adjusted mortality rate, all causes, and equation expressed as deaths per 100,000 population

Table 5.3

Correlation Matrix and Eigenvalues of X'X and X'y for the Fifteen-Factor Air Pollution Problem from McDonald & Schwing (1973) (N = 60)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
X_1	1.00														
X_2	.0922	1.00													
X_3	.5033	.3463	1.00												
X_4	.1011	-.3981	-.4340	1.00											
X_5	.2634	-.2092	.2623	-.5091	1.00										
X_6	-.4904	.1163	-.2385	-.1389	-.3951	1.00									
X_7	-.4903	.0139	-.4155	.0649	-.4095	.5515	1.00								
X_8	-.0035	-.1001	-.0610	.1620	-.1843	-.2439	.1806	1.00							
X_9	.4132	.4538	.5753	-.6378	.4194	-.2088	-.4091	-.0057	1.00						
X_{10}	-.2973	.2380	-.0214	-.1177	-.4257	.7032	.3376	-.0318	-.0044	1.00					
X_{11}	.5066	.5653	.6193	-.3098	.2599	-.4033	-.6806	-.1629	.7049	-.1852	1.00				
X_{12}	-.5318	.3508	-.3565	-.0205	-.3882	.2868	.3859	.1203	-.0259	.2037	-.1298	1.00			
X_{13}	-.4873	.3210	-.3377	-.0021	-.3584	.2244	.3476	.1653	.0184	.1600	-.1025	.9838	1.00		
X_{14}	-.1069	-.1078	-.0993	.0172	-.0041	-.2343	.1180	.4321	.1593	-.0685	-.0965	.2823	.4094	1.00	
X_{15}	-.0773	.0679	-.4528	.1124	-.1357	.1765	.1224	-.1250	-.1180	.0607	-.1522	-.0202	-.0459	-.1026	1.00
Y	.5095	-.0300	.2770	-.1746	.3573	-.5110	-.4248	.2655	.6437	-.2848	.4105	-.1772	-.0774	.4259	-.0885
λ_i^a	4.527	2.755	2.054	1.349	1.223	0.960	0.612	0.473	0.371	0.216	0.166	0.127	0.114	0.046	0.005

a. $\Sigma\lambda_i^{-1} = 263.06$

Table 5.4

Description of Variables in the Ten-Factor Gas Mileage Problem from Hocking (1976)

Variable	Description
X_1	Engine shape [straight (1) or V (0)]
X_2	Number of cylinders
X_3	Transmission type [manual (1) or automatic (0)]
X_4	Number of transmission speeds
X_5	Engine size (cubic inches)
X_6	Horsepower
X_7	Number of carburetor barrels
X_8	Final drive ratio
X_9	Weight (pounds)
X_{10}	Quarter mile time (seconds)
Y	Gas mileage (MPG)

Table 5.5

Correlations and Eigenvalues of X'X and X'y for Hocking's (1976) Gas Mileage Problem (N = 32)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	1.00									
x_2	-.811	1.00								
x_3	.168	-.523	1.00							
x_4	.206	.493	.794	1.00						
x_5	-.710	.902	-.591	-.556	1.00					
x_6	-.723	.832	-.243	-.126	.791	1.00				
x_7	-.570	.527	.058	.274	.395	.750	1.00			
x_8	.440	-.700	.713	.700	-.710	-.449	-.091	1.00		
x_9	-.555	.782	-.692	-.583	.888	.659	.428	-.712	1.00	
x_{10}	.745	-.591	.230	-.213	-.434	-.708	-.656	.091	-.175	1.00
Y	.664	-.852	.600	.480	-.848	-.776	-.551	.681	-.868	.419
λ_i^a	5.760	2.650	0.597	0.270	.222	0.210	0.133	0.081	.054	.024

a. $\Sigma\lambda_i^{-1} = 95.25$

of the overall job being done by his or her supervisor. The predictors, which were selected by management, were employee ratings on how well the supervisor (i) handled employee complaints (X_1), (ii) provided opportunities to learn new things about the job (X_2), (iii) gave raises based on performance (X_3), (iv) provided criticism of people who did a poor job (X_4), and (v) applied company policy fairly (X_5).

Each predictor was a dichotomous index. The data on 30 departments with an average of 35 employees in each was aggregated to the department level; thus, consisted of 30 observations. The matrix of zero-order correlations is presented in Table 5.6. It follows from application of Eq. (5.1) that the mean square error of the least squares estimators is $738.7\sigma^2$, which is more than 140 times what it would be for an orthogonal system.

Table 5.6 About Here

5.3 The Estimates

Eight different estimating methods were applied to each of the four models. In addition to the ridge procedure utilizing the normalization criterion, six ridge regression methods utilizing the mean square error criterion were used, and all were compared to the standard OLS procedure. The seven formulas for calculating the biasing parameter k were:

1. Normalization (Lee 1979)

$$\frac{1}{p} \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} = 1$$

Table 5.6

Correlation Matrix (X'X) and Eigenvalues for X'y for the Five-Factor Employee Satisfaction Example from Price (1977) (N = 30)

Variables	X_1	X_2	X_3	X_4	X_5
X_1	1.00				
X_2	.597	1.00			
X_3	.669	.640	1.00		
X_4	.188	.116	.377	1.00	
X_5	.867	.819	.906	.207	1.00
Y	.827	.629	.592	.158	.763
λ_i[a]	3.348	.967	.401	.283	.001

a. $\Sigma\lambda_i^{-1} = 738.7$

2. Kasarda and Shih (1977)

$$MSE = \hat{\sigma}^2 \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} + k \Sigma \frac{\hat{\alpha}_i^2}{(\lambda_i + k)^2} = \text{minimum}$$

3. Hocking, Speed and Lynn (1976)

$$k = \hat{\sigma}^2 \frac{\Sigma \lambda_i^2 \hat{\alpha}_i^2}{\Sigma \lambda_i^2 \hat{\alpha}_i^4}$$

4. Lawless and Wang (1976)

$$k = \frac{p\sigma^2}{\Sigma \lambda_i \hat{\alpha}_i^2}$$

5. Hoerl, Kennard and Baldwin (1975)

$$k = \frac{p\hat{\sigma}^2}{\Sigma \hat{\alpha}_i^2}$$

6. Hoerl and Kennard's iterative method (1976)

$$k_o = \frac{p\hat{\sigma}^2}{\Sigma \hat{\alpha}_i^2}$$

....

$$k_t = \frac{p\hat{\sigma}^2}{\Sigma \hat{\alpha}_{it}^{*2}}, \quad \hat{\sigma}_{it}^* = \frac{\lambda_i \hat{\alpha}_i}{(\lambda_i + k_t)}$$

with tolerance level $\frac{k_t + 1 - k_t}{k_t} \leq \delta$

where $\delta = 20T^{-1.30}$

$T = Tr(X'X)^{-1}/p$

7. Vinod (1976)

$$ISRM = \Sigma[(p\lambda_i/(\lambda_i + k)^2 S) - 1]^2 = \text{minimum}$$

where $S = \Sigma \lambda_i/(\lambda_i + k)^2$

Since the raw data were unavailable for data analysis purposes, the correlation matrix was used instead. This meant that the "raw" regression weights in the metric of the input variables were not available. Only standardized regression coefficients and their t-values could be reported. A t-value greater than or equal to 2.0 was considered significant at the $p \leq .05$ level.

Yet, presenting the reader with a table containing only two different statistics constitutes a formidable problem of interpretation when there are ten or more model parameters and eight different methods of estimation. Thus, a simplified table consisting of a check list of the significant parameters is provided as a way of reducing table complexity.

In the petroleum engineering example from Hoerl and Kennard (1970b) the estimates are presented in Table 5.7 and the significant parameters of the model are shown in Table 5.8. There are three obvious findings. First, the OLS results are not supported by the ridge regression results. Second, regardless of the estimating procedure used the regression of Y on X_2, X_3, X_6 and X_8, were found to be statistically significant and the regression of y on X_4 was found to be insignificant. Third, in terms of statistical significance, the results of ridge methods 4, 5, 6, and 7 were equal; and three ridge methods - the normalization, Kasarda and Shih, and Vinod's methods - departed from this norm. Only one parameter

----
Tables 5.7 and 5.8 About Here

Table 5.7

Standardized Regression Coefficients and their t-values (in parentheses) for the 10-Factor Example from Hoerl & Kennard (1970b)

Method	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1. OLS	-.166 (1.129)	-.223 (2.697)	-.360 (3.818)	-.196 (1.555)	-.479 (3.044)	.837 (4.371)	.289 (2.931)	.385 (4.166)	.082 (1.098)	.094 (0.961)
2. Normalization (Lee 1979)	-.290 (4.292)	-.129 (2.195)	-.275 (4.185)	-.064 (1.144)	-.112 (1.591)	.401 (5.993)	.094 (1.515)	.282 (4.480)	.114 (2.011)	.127 (1.917)
3. Kasarda & Shih (1977)	-.189 (1.368)	-.212 (2.638)	-.354 (3.839)	-.102 (1.509)	-.436 (2.953)	0.778 (4.448)	.267 (2.817)	.377 (4.188)	.087 (1.190)	.100 (1.041)
4. Hocking, Speed & Lynn (1976)	-.236 (2.043)	-.187 (2.493)	-.336 (3.901)	-.091 (1.397)	-.331 (2.682)	.643 (4.706)	.214 (2.499)	.354 (4.259)	.099 (1.432)	.112 (1.255)
5. Hoerl, Kennard & Baldwin (1975)	-.234 (2.013)	-.188 (2.499)	-.337 (3.898)	-.091 (1.402)	-.335 (2.694)	.648 (4.693)	.216 (2.513)	.355 (4.256)	.098 (1.422)	.111 (1.246)
6. Lawless & Wang (1976)	-.253 (2.417)	-.174 (2.426)	-.326 (3.938)	-.085 (1.344)	-.283 (2.522)	.584 (4.874)	.188 (2.327)	.342 (4.301)	.103 (1.553)	.116 (1.370)
7. Hoerl & Kennard's Iterative (1976)	-.234 (2.013)	-.188 (2.499)	-.337 (3.898)	-.091 (1.402)	-.335 (2.694)	.648 (4.693)	.216 (2.513)	.355 (4.256)	.098 (1.422)	.111 (1.246)
8. Vinod (1976)	-.270 (7.087)	-.085 (2.007)	-.210 (4.850)	-.043 (.985)	.016 (.440)	.280 (8.360)	.019 (.466)	.191 (4.496)	.100 (2.332)	.118 (2.760)

Table 5.8

The Significant Parameters (✓) in the Hoerl & Kennard 10-Factor Example by Estimating Method

Method	Parameters									
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1. OLS		✓	✓		✓	✓	✓	✓		
2. Normalization (Lee, 1979)	✓	✓	✓			✓		✓	✓	
3. Kasarda & Shih (1977)		✓	✓		✓	✓	✓	✓		
4. Hocking, Speed & Lynn (1976)	✓	✓	✓		✓	✓	✓	✓		
5. Hoerl, Kennard & Baldwin (1975)	✓	✓	✓		✓	✓	✓	✓		
6. Lawless & Wang (1976)	✓	✓	✓		✓	✓	✓	✓		
7. Hoerl & Kennard's Iterative (1976)	✓	✓	✓		✓	✓	✓	✓		
8. Vinod (1976)	✓	✓	✓			✓		✓	✓	✓

in the Kasarda and Shih results departed from the norm compared to the three discrepancies for the normalization method and four for the Vinod method.

In the environmental sciences model of air pollution and mortality rates the complicating factor was the sheer mass of data for the fifteen-factor model. The estimates for this problem are presented in Table 5.9. A summary table of the significant air pollution model parameters is also provided in Table 5.10. Again, most of the ridge estimates depart from OLS estimates; and, as well, the ridge estimates differ among themselves. The Vinod method is the most deviant and the Kasarda and Shih result is the same as the OLS result. Methods 2, 4, 5, 6 and 7 generated identical conclusions in that the regression of Y on X_1, X_2, X_9 and X_{14} was found to be significant, while the regression of Y on X_3, X_4, X_5, X_7, X_{10} thru X_{13}, and X_{15} was found to be insignificant. Perhaps the most interesting finding in the McDonald and Schwing example is the fact that the estimates from the normalization ridge regression analysis on the one hand and from Hoerl and Kennard's iterative ridge regression on the other, were almost identical. In fact, eleven of the fifteen parameter estimates were the same to two places of decimals.

Tables 5.9 and 5.10 About Here

Inconsistencies in the patterning of the standardized regression estimates for the gas mileage problem and the employee satisfaction problem were also noticed. For example,

Table 5.9

Standardized Regression Coefficients and their t-values (in parentheses) for the 15-Factor Air Pollution Problem from McDonald & Schwing (1973)

	Method	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
1.	OLS	.297 (2.03)	-.307 (1.72)	-.229 (1.59)	-.200 (1.02)	-.228 (1.52)	-.231 (1.46)	-.060 (0.42)	-.083 (0.89)	.651 (3.48)	-.018 (0.15)	-.029 (0.14)	-.997 (1.14)	1.000 (1.34)	.087 (0.58)	-.018 (0.10)
2.	Normalization (Lee 1979)	.256 (3.35)	-.189 (2.56)	-.103 (1.28)	-.063 (0.88)	-.031 (0.40)	-.118 (1.47)	-.095 (1.24)	.120 (1.73)	.464 (5.90)	-.036 (0.49)	.034 (0.44)	-.055 (1.08)	.051 (1.06)	.245 (3.47)	.032 (0.47)
3.	Kasarda & Shih (1977)	.303 (2.14)	-.301 (1.77)	-.223 (1.58)	-.175 (0.95)	-.210 (1.45)	-.203 (1.35)	-.070 (0.50)	.086 (0.93)	.650 (3.62)	-.031 (0.27)	-.018 (0.09)	-.623 (1.39)	.615 (1.33)	.147 (1.24)	.008 (0.08)
4.	Hocking, Speed & Lynn (1976)	.280 (2.95)	-.231 (2.37)	-.146 (1.43)	-.088 (0.89)	-.086 (0.88)	-.131 (1.27)	-.092 (0.96)	.110 (1.41)	.542 (5.11)	-.041 (0.46)	.019 (0.17)	-.087 (1.23)	.078 (1.15)	.241 (2.97)	.024 (0.30)
5.	Hoerl, Kennard & Baldwin (1975)	.297 (2.56)	-.269 (2.08)	-.187 (1.52)	-.120 (0.89)	-.145 (1.20)	-.154 (1.21)	-.086 (0.74)	.098 (1.14)	.607 (4.36)	-.044 (0.42)	0.00 (0.02)	-.170 (1.38)	.155 (1.27)	.224 (2.46)	.013 (0.15)
6.	Lawless & Wang (1976)	.269 (3.15)	-.211 (2.48)	-.125 (1.36)	-.075 (0.88)	-.057 (0.66)	-.124 (1.35)	-.094 (1.09)	.116 (1.56)	.504 (5.50)	-.039 (0.47)	.027 (0.29)	-.068 (1.15)	.062 (1.10)	.244 (3.21)	.028 (0.39)
7.	Hoerl & Kennard's Iterative (1976)	.253 (3.40)	-.184 (2.58)	-.098 (1.26)	-.061 (0.88)	-.025 (0.33)	-.117 (1.51)	-.095 (1.27)	.125 (1.77)	.454 (6.00)	-.036 (0.50)	.036 (0.48)	-.052 (1.07)	.049 (1.05)	.245 (3.54)	.033 (0.49)
8.	Vinod (1976)	.192 (4.20)	-.106 (2.51)	-.031 (0.69)	-.038 (0.93)	.039 (0.89)	-.109 (2.54)	-.091 (1.99)	.118 (2.47)	.309 (7.49)	-.036 (0.78)	.056 (1.44)	-.031 (1.00)	.027 (0.90)	.216 (4.57)	.030 (0.64)

Table 5.10

The Significant Parameters (✓) in the McDonald and Schwing 15-Factor Air Pollution Problem by Estimating Method

Method	Parameters														
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
1. OLS	✓								✓						
2. Normalization (Lee, 1979)	✓	✓							✓					✓	
3. Kasarda & Shih (1977)	✓								✓						
4. Hocking, Speed & Lynn (1976)	✓	✓							✓					✓	
5. Hoerl, Kennard & Baldwin (1975)	✓	✓							✓					✓	
6. Lawless & Wang (1976)	✓	✓							✓					✓	
7. Hoerl and Kennard's Iterative (1976)	✓	✓							✓					✓	
8. Vinod	✓	✓				✓		✓	✓					✓	

in the gas mileage model the OLS result was reinforced by the findings from four of the ridge regression analyses. While the results of the normalization, Hoerl and Kennard's iterative, and Vinod's methods differed from the norm, two of these - Hoerl and Kennard's iterative and Vinod's - were the same. Both found parameters p_{y,x_2}, p_{y,x_3}, p_{y,x_6}, p_{y,x_7}, and p_{y,x_9} to be significant. The normalization results were different from all of these in that two parameters p_{y,x_7} and p_{y,x_9} were found significant.

If there is consensus concerning which parameters are significant among the seven ridge regression procedures, does this mean that deviant results are probably erroneous? If the fourth example is anything to go by the answer is in the negative. The interesting finding in the employee satisfaction study was that while the Kasarda and Shih, and Hoerl, Kennard and Baldwin methods generated the same results the standardized coefficients of the significant parameters were greater than unity; hence, uninterpretable. It is also worth noting that in this example, as was the case in the second example, the normalization results and the Hoerl and Kennard iterative results were very close to one another.

Tables 5.11, 5.12, 5.13 and 5.14 About Here

5.4 Performance Comparisons

In terms of performance there is little that can be concluded from Tables 5.7 through 5.14. In general, Kasarda

Table 5.11

Standardized Regression Coefficients and their t-values (in parentheses) for the 10-Factor Gas Mileage Problem from Hocking (1976)

Estimating Method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1. OLS	.023 (.137)	-.034 (.112)	.210 (1.267)	.078 (.440)	.272 (.756)	-.239 (.991)	-.053 (.245)	.069 (.487)	-.605 (2.013)	.248 (1.172)
2. Normalization (Lee , 1979)	.049 (.605)	.093 (1.280)	.161 (1.987)	.071 (.889)	-.081 (1.149)	-.139 (1.776)	-.174 (2.336)	.089 (1.075)	-.257 (3.503)	.067 (.886)
3. Kasarda & Shih (1977)	.033 (.266)	-.058 (.371)	.187 (1.487)	.080 (.629)	.003 (.025)	-.155 (1.105)	-.168 (.146)	.083 (.719)	-.358 (2.683)	.125 (.956)
4. Hocking, Speed & Lynn (1976)	.032 (.245)	-.054 (.320)	.189 (1.455)	.081 (.610)	.018 (.121)	-.159 (1.070)	-.163 (1.351)	-.083 (.695)	-.374 (2.613)	.134 (.972)
5. Hoerl, Kennard & Baldwin (1975)	.041 (.396)	-.077 (.712)	.174 (1.687)	.075 (.731)	-.048 (.491)	-.145 (1.355)	-.178 (1.916)	.086 (.858)	-.300 (3.041)	.090 (.898)
6. Lawless & Wang (1976)	.037 (.320)	-.067 (.509)	.181 (1.571)	.078 (.672)	-.024 (.204)	-.149 (1.205)	-.175 (1.681)	.085 (.778)	-.328 (2.844)	.107 (.925)
7. Hoerl & Kennard's Iterative (1976)	.057 (.906)	-.104 (2.08)	.148 (2.373)	.068 (1.117)	-.099 (1.933)	-.136 (2.350)	-.162 (2.717)	.092 (1.386)	-.226 (4.067)	.054 (.948)
8. Vinod (1976)	.062 (1.141)	-.108 (2.691)	.140 (2.645)	.067 (1.294)	-.106 (2.483)	-.133 (2.769)	-.153 (2.938)	.093 (1.629)	-.210 (4.463)	.050 (1.028)

Table 5.12

The Significant Parameters (✓) in the Hocking 10-Factor Gas Mileage Problem by Estimating Method

Estimating Method	Parameters									
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1. OLS									✓	
2. Normalization							✓		✓	
3. Kasarda & Shih									✓	
4. Hocking, Speed & Lynn									✓	
5. Hoerl, Kennard & Baldwin									✓	
6. Lawless & Wang									✓	
7. Hoerl & Kennard's Iterative		✓	✓			✓	✓		✓	
8. Vinod		✓	✓		✓	✓	✓		✓	

Table 5.13

Standardized Regression Coefficients and their t-values (in parentheses) for Price's (1977) Employee Satisfaction Problem

		X_1	X_2	X_3	X_4	X_5
1.	OLS	2.492 (3.27)	1.537 (2.70)	2.370 (2.32)	-.405 (1.98)	-4.720 (2.37)
2.	Normalization	.569 (5.20)	.181 (1.64)	-.022 (.204)	.011 (.119)	.128 (2.606)
3.	Kasarda & Shih	1.775 (3.695	1.008 (2.751)	1.403 (2.197)	-.237 (1.579)	-2.820 (2.306)
4.	Hocking, Speed & Lynn	.627 (5.201)	.200 (1.657)	-.019 (.156)	.010 (.099)	.067 (.975)
5.	Hoerl, Kennard & Baldwin	1.66 (3.802)	.926 (2.757)	1.252 (2.162)	-.211 (1.479)	-2.525 (2.287)
6.	Lawless & Wang	.612 (5.193)	.194 (1.642)	-.022 (.188)	.010 (.108)	.086 (1.386)
7.	Hoerl & Kennard's Iterative	.520 (5.281)	.171 (1.709)	-.010 (.102)	.010 (.116)	.156 (3.855)
8.	Vinod	.691 (5.273)	.231 (1.786)	.015 (.108)	.003 (.035)	-.038 (.358)

Table 5.14

The Significant Parameters (✓) in the Price 5-Factor Employee Satisfaction Study by Estimating Method

Estimating Method	Significant Parameters				
	X_1	X_2	X_3	X_4	X_5
1. OLS	✓	✓	✓		✓
2. Normalization	✓				✓
3. Kasarda & Shih	✓	✓	✓		✓
4. Hocking, Speed & Lynn	✓				
5. Hoerl, Kennard & Baldwin	✓	✓	✓		✓
6. Lawless & Wang	✓				
7. Hoerl & Kennard's Iterative	✓				✓
8. Vinod	✓				

and Shih's method seemed "conservative" while Vinod's method seemed inconsistent. Thus, there were few differences between OLS and Kasarda and Shih's results; whereas, in contrast, Vinod's results had little in common with the results generated by any other method. As expected, OLS estimates were inflated and the OLS t-values indicated that the standard errors were excessively high in comparison with the ridge regression estimates. But even if OLS, Kasarda and Shih, and Vinod methods are rejected how does one choose between the remaining methods? One answer, albeit a rather subjective one, lies in the making of comparisons between the performance indices of each method. See sections 4.4.1 to 4.4.8 above.

Because estimation bias is dependent on unknown population parameters it is not possible to use a single performance measure such as MSE or Eft (see section 2.2.14) or any other single method which involves bias introduced by the ridge procedure. Instead several indices of performance have to be evaluated. In general, for any ridge regression estimation method it is desirable:

(i) to reduce the maximum variance inflation factor (V_{max}) to a value close to unity (Note that the criterion used by the normalization method is normalizing the average variance inflation factor to unity);

(ii) to generate the maximum reduction of the relative degree of multicollinearity of the problem (D_{max}) without introducing too large a bias (k) (That is, the closer the V_{max} is to unity and D_{max} to zero, the better the ridge performance);

(iii) to minimize the empirical sensitivity ratio (E.S. Ratio) or sampling variance;

(iv) to reduce the total variance in the OLS estimates inflated by the presence of multicollinearity (But note that RVAR is the relative reduction in the total variance in the OLS estimates; and, therefore, should not be used alone to judge the ridge performance, because the larger the RVAR the larger the bias);

(v) to produce low, or other things being equal, acceptably low inflation in the residual sum of squares (ϕ_o);

(vi) to generate the lowest measure of the average percentage error of estimated variance (Av.%Er.VAR) that has not been "eliminated" (including suppressed variance) by the respective ridge procedure (Note that other things being equal the Av.%Er.VAR index is probably the best indicator of ridge performance).

If a set of ridge estimates satisfactorily meets these performance criteria and is found α-unacceptable; that is, where the results are significantly different from the OLS estimates, the intuitive evidence would support rejection of the OLS procedure. For an alternative view see McCabe (1978). The analyst is also interested in an estimating procedure which performs consistently from one data set to the next; that is, one which consistently reduces V_{max} to a value close to unity; which consistently generates a low sensitivity ratio, and so on.

The results of the eight analyses for the four examples are presented in Tables 5.15 through 5.18. From the performance indices for the Kasarda and Shih and Vinod methods it is obvious that the tentative conclusions drawn from an examination of the parameter estimates in Tables 5.7 to 5.14 were supported. Thus, the Kasarda and Shih method consistently failed to significantly reduce the variance inflation factor (V_{max}) and compared to the other methods it generated little reduction in the relative degree of multicollinearity (D_{max}). Likewise it was consistently the least effective method for reducing both the sampling variance (E.S. Ratio) and the average percentage error of estimated variance (Av.%Er.VAR). In short, the method was little different than OLS regression and consequently reflected most of the limitations of the OLS procedure in the presence of multicollinearity.

Tables 5.15, 5.16, 5.17, and 5.18 About Here

In some respects Vinod's method seems a good one -- compare the RVAR and E.S. Ratio statistics for Vinod with those for the other six ridge regression methods. If, however, we look at the size of its biasing parameter k, its negative D_{max}, and its low (less than unity) V_{max} value, it can be shown that Vinod's method consistently oversuppressed the estimates and definitely produced too high a bias. Furthermore, it sometimes produced more than one set of ridge estimates as in the Price example (see the five-factor problem, Table 5.18), although theoretically there should be only one

Table 5.15

The Characteristics and Performance Indices of OLS and Ridge Regression Procedures for the 10-Factor Example from Hoerl & Kennard (1970b)

($p = 10$, $N = 36$, $D_c = .952$)

Estimating Method	k	V_{max}	D_{max}	E.S. Ratio	RVAR	ϕ_o	α	Av.%Er.VAR	R^2
1. OLS	0	9.10	.922	1.000	0	0	1.00	159.2%	.895
2. Normalization	.150	1.23	.145	2.94×10^{-2}	70.56%	37.0%	.498	12.64%	.715
3. Kasarda & Shih	.007	7.59	.904	.655	10.60%	.49%	1.00	204.14%	.877
4. Hocking et al.	.030	4.63	.829	.230	33.39%	5.66%	.998	126.58%	.834
5. HKB	.029	4.73	.833	.240	32.55%	5.34%	.999	129.44%	.835
6. Lawless & Wang	.045	3.56	.763	.143	42.74%	9.96%	.985	104.48%	.812
7. H & K's Iterative	.029	4.73	.833	.240	32.55%	5.34%	.999	129.44%	.835
8. Vinod	.458	0.46	-.313	5.07×10^{-3}	87.83%	89.2%	.042	58.59%	.559

Table 5.16

The Characteristics and Performance Indices of OLS and Ridge Regression Procedures for the 15-Factor Air Pollution Problem from McDonald & Schwing (1973)

($p = 15$, $N = 60$, $D_c = .997$)

	Estimating Method	k	V_{max}	D_{max}	E.S. Ratio	RVAR	ϕ_o	α	Av.%Er.VAR	R^2
1.	OLS	0	103.20	.994	1.000	0	0	1.000	1689.8%	.765
2.	Normalization	.145	1.24	.151	3.69×10^{-6}	94.43%	13.0%	.975	17.4%	.607
3.	Kasarda & Shih	.003	40.79	.984	2.31	50.11%	.61%	1.000	794.4%	.754
4.	Hocking et al.	.070	2.33	.589	4.18×10^{-5}	90.86%	7.05%	.999	62.1%	.667
5.	HKB	.028	4.37	.817	1.32×10^{-3}	84.71%	3.72%	1.00	174.1%	.715
6.	Lawless & Wang	.102	1.66	.372	1.15×10^{-4}	92.85%	9.62%	.994	38.0%	.639
7.	H & K's Iterative	.157	1.16	.104	2.86×10^{-6}	94.76%	14.0%	.966	15.0%	.599
8.	Vinod	.488	0.44	-.326	9.81×10^{-8}	98.06%	36.9%	.378	65.2%	.449

Table 5.17

The Characteristics and Performance Indices of OLS and Ridge Regression Procedures for the 10-Factor Gas Mileage Problem from Hocking (1976)

($p = 10$, $N = 32$, $D_c = .985$)

	Estimating Method	k	V_{max}	D_{max}	E.S. Ratio	RVAR	ϕ_o	α	Av.%Er.VAR	R^2
1.	OLS	0	21.79	.969	1.000	0	0	1.000	861.2%	.869
2.	Normalization	.174	1.16	.102	3.68×10^{-3}	89.59%	8.79%	.995	9.6%	.796
3.	Kasarda & Shih	.049	4.12	.802	7.90×10^{-3}	69.67%	3.75%	1.000	19.1%	.840
4.	Hocking et al.	.042	4.77	.835	.011	66.14%	3.26%	1.000	22.5%	.843
5.	HKB	.096	1.98	.494	.491	81.91%	6.09%	.999	7.39%	.822
6.	Lawless & Wang	.068	2.94	.697	5.53×10^{-3}	76.17%	4.83%	1.000	12.9%	.832
7.	H & K's Iterative	.291	.737	-.164	2.22×10^{-3}	93.99%	12.15%	.982	42.2%	.762
8.	Vinod	.386	.546	-.271	1.54×10^{-3}	95.68%	14.77%	.965	58.4%	.736

Table 5.18

The Characteristics and Performance Indices of OLS and Ridge Regression Procedures for the 5-Factor Employee Satisfaction Study (Price, 1977) by Estimating Method

($p = 5$, $N = 30$, $D_c = .985$)

	Estimating Method	k	V_{max}	D_{max}	E.S. Ratio	RVAR	ϕ_o	α	Av.%Er.VAR	R^2
1.	OLS	0	421.8	.998	1.00	0	0	1.000	12590.6%	.766
2.	Normalization	.092	1.30	.186	1.41×10^{-6}	99.21%	26.02%	.295	33.0%	.636
3.	Kasarda & Shih	.001	159.4	.996	.534	61.52%	3.65%	.967	956.8%	.729
4.	Hocking et al.	.053	1.59	.339	1.64×10^{-6}	99.02%	24.07%	.336	44.0%	.663
5.	HKB	.0013	129.9	.995	.487	68.44%	4.87%	.939	781.0%	.724
6.	Lawless & Wang	.061	1.50	.295	1.86×10^{-6}	99.07%	24.51%	.327	41.4%	.657
7.	H & K's Iterative	.141	1.07	.045	6.98×10^{-7}	99.36%	28.38%	.252	22.7%	.607
8.	Vinod (#1)	.030	1.99	.498	2.50×10^{-6}	98.75%	22.48%	.373	11.68%	.681
	Vinod (#2)	.964	.250	-.410	2.39×10^{-8}	99.87%	61.90%	.026	83.2%	.369

set of estimates as noted in section 3.3 above.

In the next section (5.5) where the ridge trace results are presented (Tables 5.1 through 5.4) the locations on the ridge traces show that neither the Kasarda and Shih nor the Vinod methods satisfy Hoerl and Kennard's guidelines for a desirable estimator (Hoerl and Kennard, 1970a). It is for these reasons that the Kasarda and Shih, and Vinod methods were dropped from further consideration. With regard to the four examples neither method proved satisfactory in terms of the performance indices compared to the other five ridge regression methods.

From the performance indices of the remaining five ridge methods it is obvious that the normalization method and Hoerl and Kennard's iterative method are the best. In Table 5.19 the data in Tables 5.15 through 5.18 has been collated in a simplified form. The rankings for each performance index for each of the five remaining methods for each of the four examples are presented. Consider the D_{max} rankings in the Hoerl and Kennard example (Table 5.19). The lowest degree of multicollinearity reported is for method two, the normalization method [D_{max} (Table 5.15) = .145] which is ranked "1"; the next lowest is for method 6, Lawless and Wang's method (D_{max} = .763) which is ranked "2" and so on. The lowest ranked methods are for the HKB and H & K iterative; both of which have the same D_{max} values (.833) and which rank 4th-equal.

Table 5.19 About Here

Table 5.19

Rankings of Performance Indices for Five Ridge Regression Methods*

Performance	Hoerl & Kennard Example					McDonald & Schwing Example					Hocking Example					Price Example				
	Methods					Methods					Methods					Methods				
Criteria	2	4	5	6	7	2	4	5	6	7	2	4	5	6	7	2	4	5	6	7
k^a	5	3	1	4	1	4	2	1	3	5	4	1	3	2	5	4	2	1	3	5
$V_{max}{}^b$	1	3	4	2	4	2	4	5	3	1	2	5	3	4	1	2	4	5	3	1
$D_{max}{}^c$	1	3	4	2	4	2	4	5	3	1	1	5	3	4	2	2	4	5	3	1
E.S. Ratio[d]	1	3	4	2	4	2	3	5	4	1	2	4	5	3	1	2	3	5	4	1
RVAR[e]	1	3	4	2	4	2	4	5	3	1	2	5	3	4	1	2	4	5	3	1
$\phi_o{}^f$	1	3	4	2	4	2	4	5	3	1	2	5	3	4	1	2	4	5	3	1
α^g	Rej	Acc	Acc	Rej	Acc	Rej	Acc	Acc	Rej	Rej	Acc	Acc	Acc	Acc	Rej	Rej	Rej	Rej	Rej	Rej
Av.%Er.VAR[h]	1	3	4	2	4	2	4	5	3	1	2	4	1	3	5	2	4	5	3	1
$R^{2\,i}$	5	3	1	4	1	4	2	1	3	5	4	1	2	3	5	4	2	1	3	5

a. Ranked from lowest k (1) to highest k-value (5).

b. Ranked from lowest VIF_{max} (1) to highest VIF_{max} (5).

c. Ranked from lowest degree of multicollinearity to highest.

d. Ranked from least sensitive to sampling error relative to OLS, to most sensitive.

e. Ranked from highest percentage reduction in variance inflation to the lowest.

f. Ranked from the lowest percentage of inflation in the residual sum of squares to the highest.

g. Whether the ridge estimate is α-acceptable or not; that is, whether it falls in the $(1 - \alpha)$ 100% confidence region of the OLS estimates. (Acc = α-acceptable; Rej = reject or α-unacceptable).

h. Ranked from lowest average percentage error variance (1) to highest (5).

i. Ranked from highest R^2 (1) to lowest (5).

*Method 2 = normalization ridge regression
Method 4 = Hocking, Speed & Lynn (1976)
Method 5 = Hoerl, Kennard & Baldwin (1975)
Method 6 = Lawless & Wang (1976)
Method 7 = Hoerl & Kennard (1976)

Hoerl and Kennard's iterative method outranks the normalization method for three examples (examples two, three and four) but not for the first. For the Hoerl and Kennard (1970b) example the Hoerl and Kennard iterative results are poor. The variance is inflated by a factor of 4.73 and the average percentage error of estimated variance is higher than that of the normalization method by a factor greater than ten. Obviously, in this example the iterative method has not maximized the reduction of the harmful effects of multicollinearity as compared to the normalization method.

Though it seems as if the iterative method may not be consistent it is likely that the iterative results were artifactual; hence, probably correctable. The results for example one were obtained from an iterative procedure which ended after a single step. In other words, the iterative procedure was not in operation. Though the multicollinearity in the Hoerl and Kennard example was high ($D_C = .952$) it was the least severe of the four examples. It might be that the iterative method is ineffective except for problems with extremely severe multicollinearity. For problems with severe collinearity, however, the iterative performance compares most favourably with that of the normalization method and more than favourably with the other methods. Modification to the tolerance levels of the iterative method could resolve the problem - though computer time would no doubt be increased.

Note that a weakness of the iterative method might be its tendency to oversuppress the estimates as in example number three, the gas mileage problem, where the $D_{max} = -0.164$. In general, estimation methods which overcorrect the multicollinearity problem can be regarded as unacceptable.

Perhaps the chief advantage, then, of the normalization method lies not in its superiority -- in three examples it is outperformed by the H & K iterative method -- but by its consistency.

(i) It consistently reduced the variance inflation factor to a value close to unity.

(ii) It consistently generated considerable reduction in the relative degree of multicollinearity of the problem and did so without introducing too large a bias.

(iii) It consistently had a very low sensitivity ratio; that is, always produced more stable estimates than those produced by OLS.

(iv) It consistently produced a large amount of reduction in the total variance of the estimates.

(v) It generally produced low or acceptably low inflation in the residual sum of squares.

(vi) It consistently produced low average percentage error in the variance of the estimated coefficients (Av.%Er.VAR), though it did not completely eliminate the variance inflation caused by multicollinearity.

(vii) It was usually significantly different from the OLS estimation procedure.

(viii) As is shown in section 5.5 it consistently produced a k-value which satisfied Hoerl and Kennard's subjective guidelines for a desirable estimator (Hoerl and Kennard, 1970a).

5.5 The Ridge Trace Results

In this section the ridge traces of the four examples are plotted. The purpose of each trace is not to try to determine the optimal k, because the method as explained in section 3.1 is much too subjective and arbitrary. Rather, each trace is used to display the characteristics of the underlying data set while at the same time facilitating comparisons between the k's obtained from the seven different ridge regression procedures. The ridge traces for each sample are given in Figures 5.1 through 5.4).

Figures 5.1, 5.2, 5.3 and 5.4 About Here

It is obvious from the ridge traces of the four examples that the normalization method is the only one that consistently produces estimates which satisfy the Hoerl and Kennard (1970a) guidelines listed in section 3.1.

5.6 Conclusions

The purpose of this chapter was to examine the operation of ridge regression estimation in practice. This called for the making of three kinds of comparisons; namely, comparisons between parameter estimates, comparisons between performance indices, and comparisons between k-values for different ridge regression estimation methods on the ridge traces. Little could be learned from the parameter estimates comparisons except that Kasarda and Shih's method seemed too "conservative" and Vinod's method for several reasons too

Fig. 5.1 Ridge trace result for the 10-factor example from Hoerl & Kennard (1970b)

(N=36, V_{max}=9.11, D_{max}=.952)

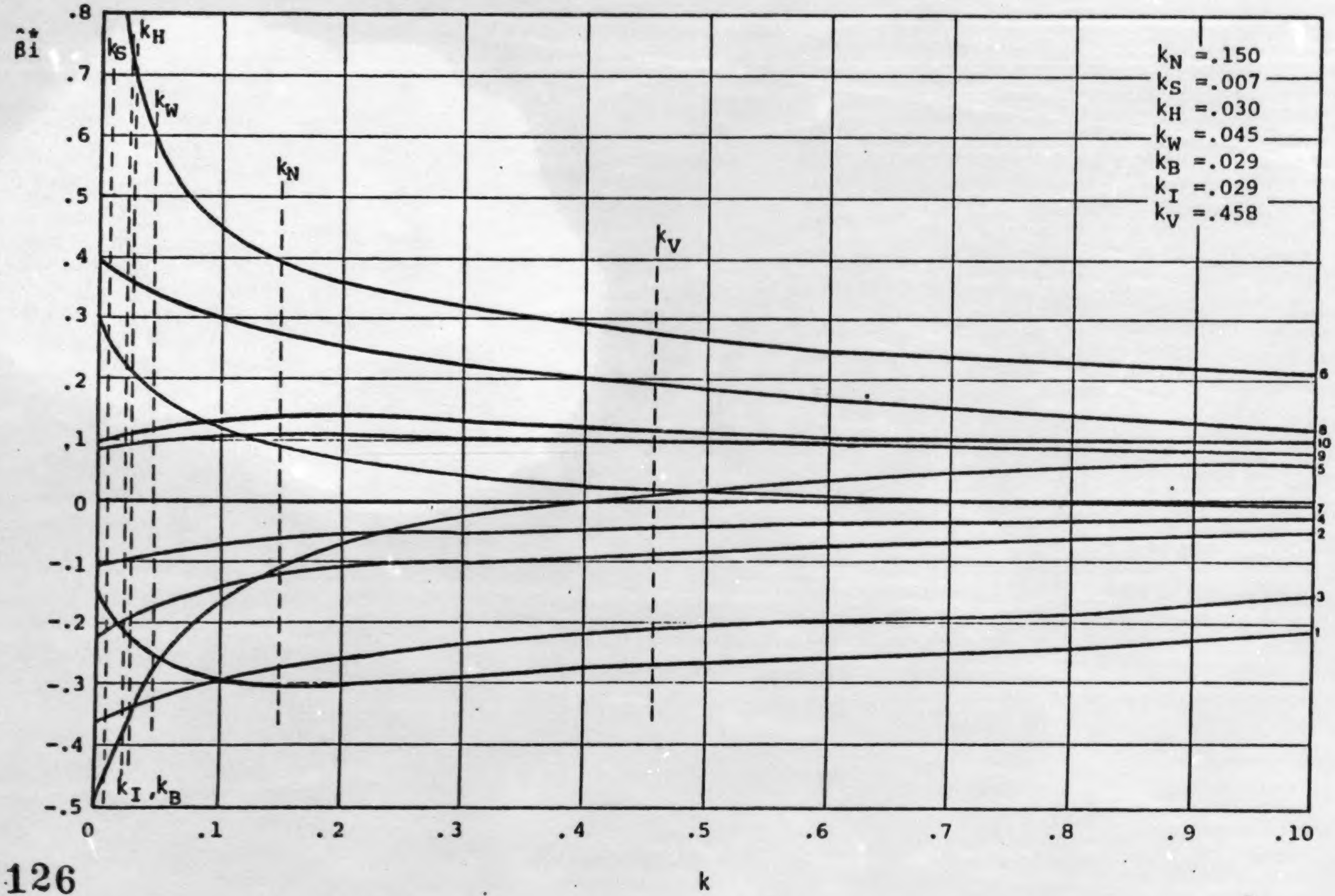


Fig. 5.2 Ridge Trace Result for the Air Pollution Problem from McDonald & Schwing (1973)

(N=60, V_{max}=104.2 D_{max} =.994)

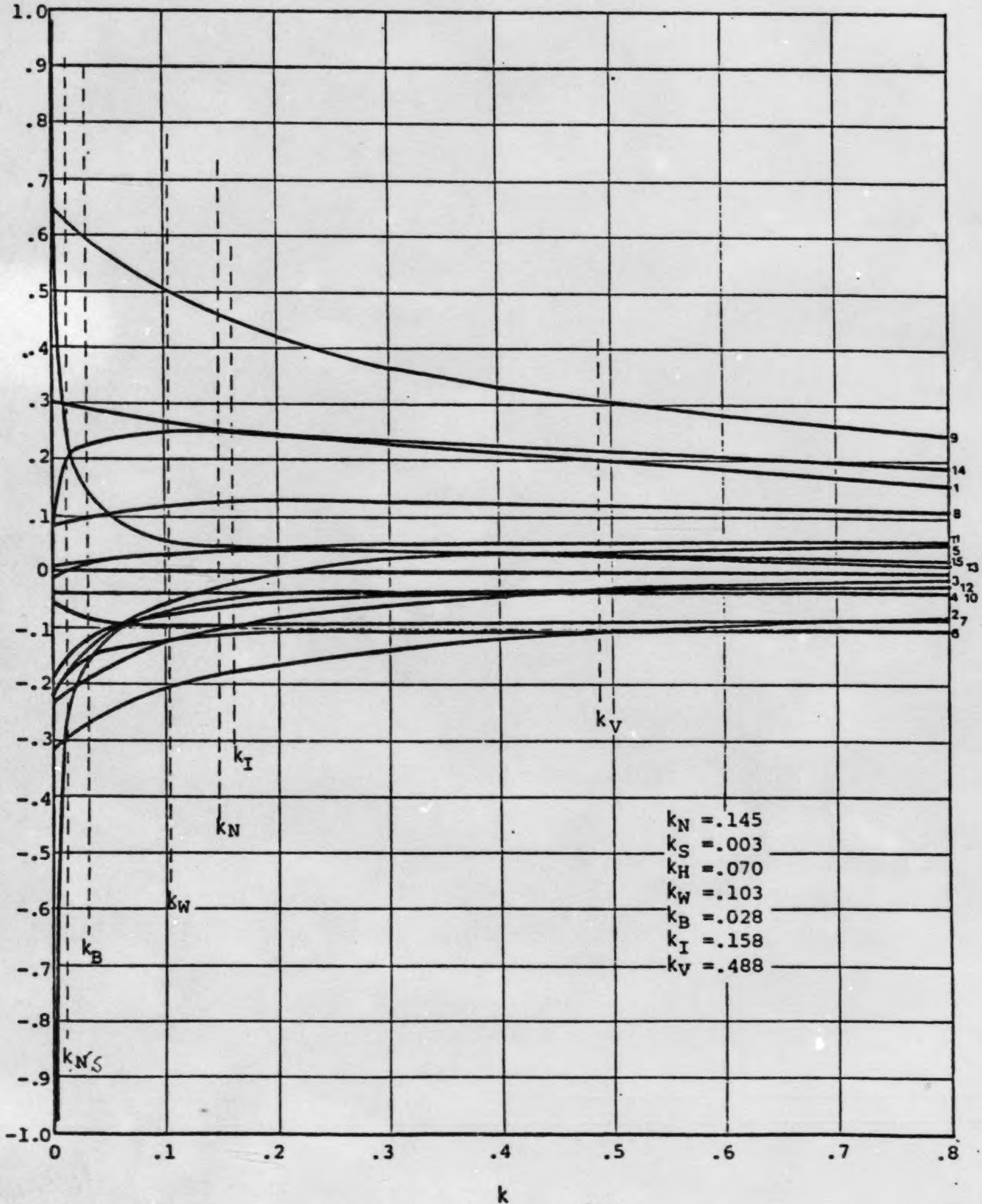
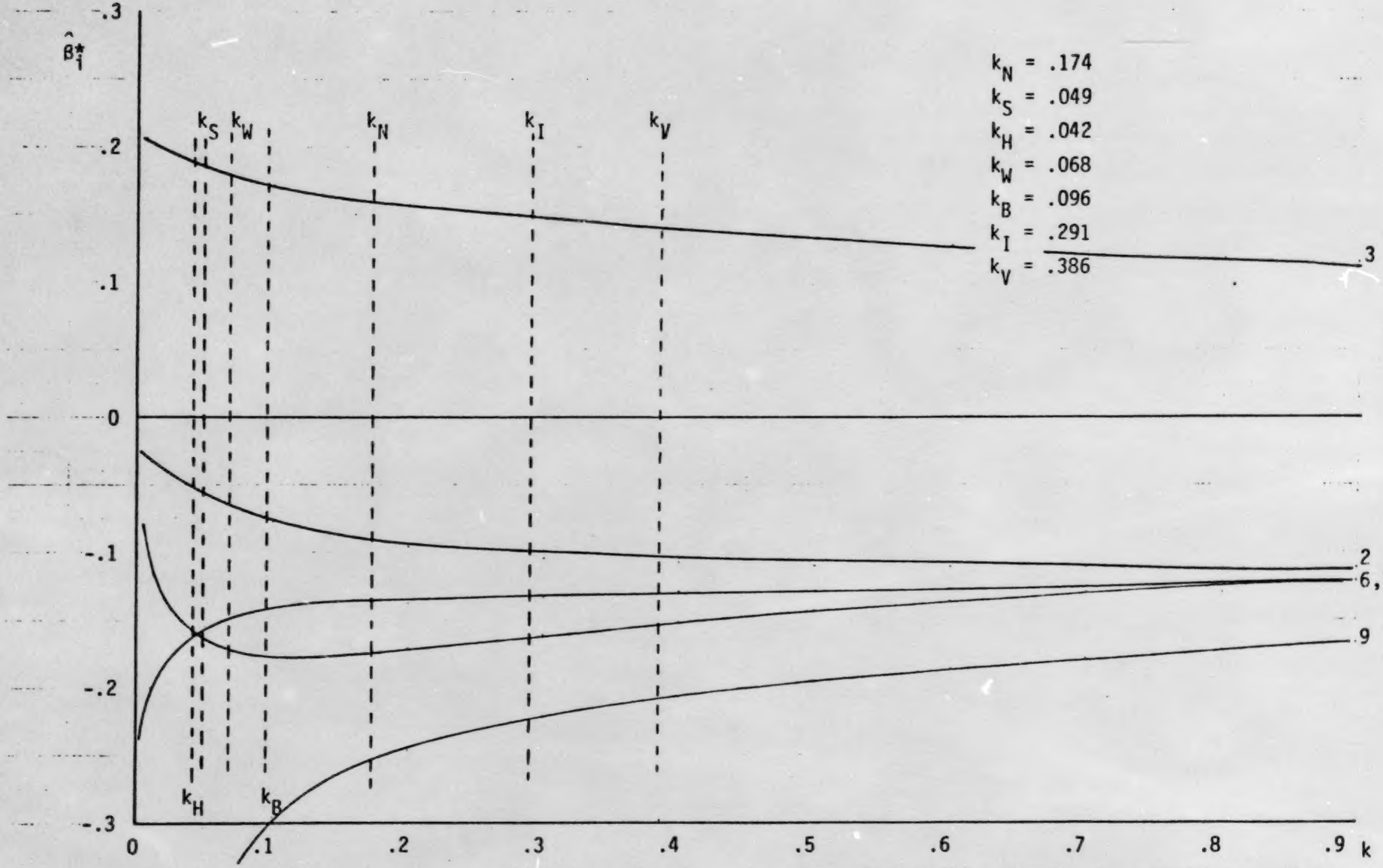


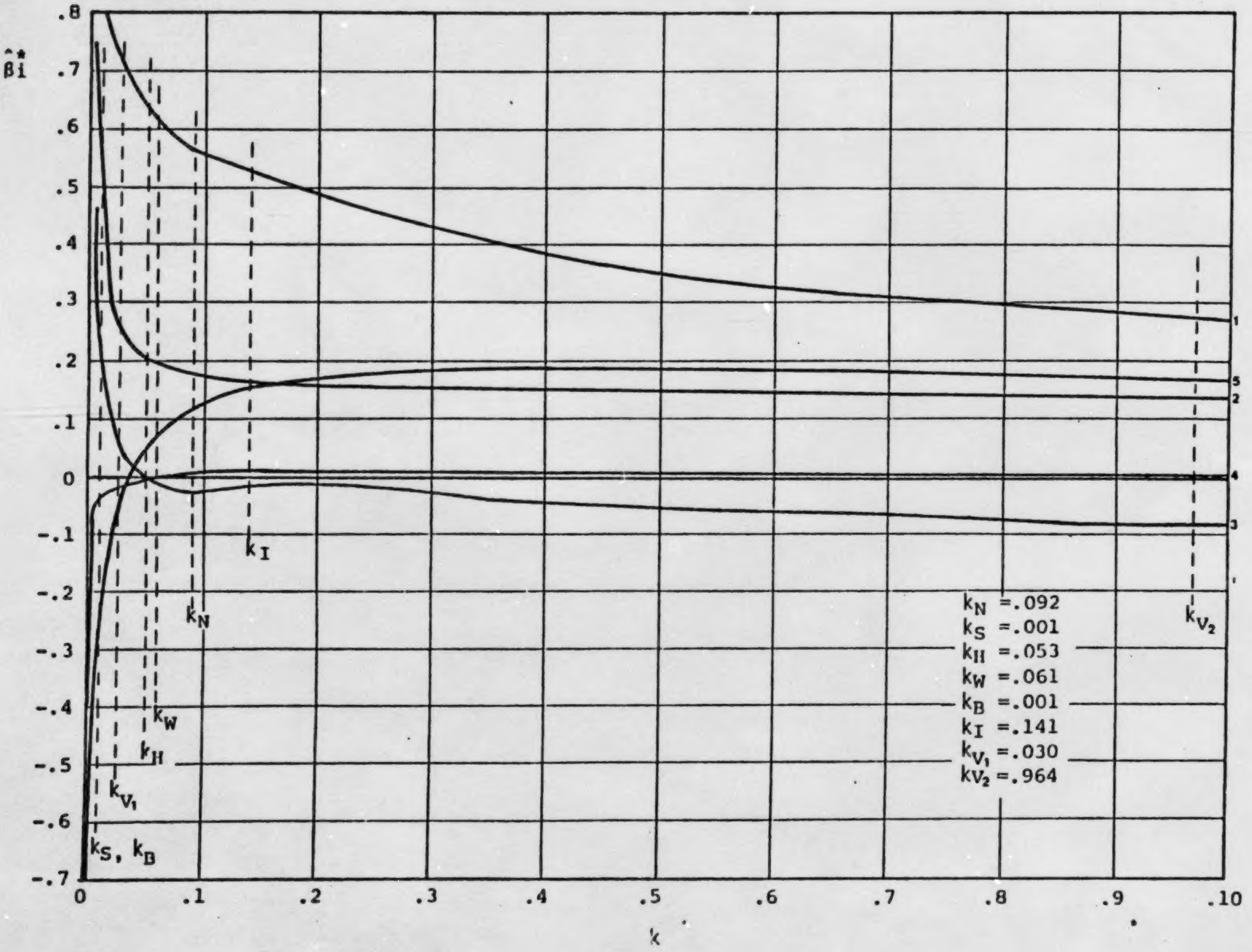
Figure 5.3. Ridge Trace of the 10-Factor Gas Mileage Problem from Hocking (1976)
(N = 32, Vmax = 21.79, Dc = .985)



Where: k_N = estimated k using normalization criterion; k_S = Kasarda & Shih's method; k_H = Hocking et al.; k_W = Lawless & Wang; k_B = HKB; k_I = HK's Iterative; and k_V = Vinod's method.

Fig. 5.4 Ridge Trace Result for the 5-factor Employee Satisfaction Problem from Price (1977)

($N=30$, $V_{max}=421.8$ $D_{max}=.998$)



inconsistent. Since the comparisons between the performance indices confirmed this tentative conclusion, the two methods were dropped from further consideration.

Of the five remaining methods, the normalization method and Hoerl and Kennard's iterative seemed to be the best. It is important to note, however, that the comparisons on the basis of performance indices does not constitute proof of performance superiority. The comparative method falls short of assuring judgmental certainty. Thus, at best we can only conclude that in terms of performance the normalization method seems to perform satisfactorily; and, further, seems to be a consistent performer. Many more examples of ill-conditioned data sets have to be estimated before the analyst can confidently claim performance superiority for one ridge procedure or the other. On the basis of the available evidence from four practical examples we conclude only that the normalization method seems a viable one -- one to be taken as seriously as other ridge methods. The ridge trace results confirmed this tentative conclusion.

Thus, despite our obvious vested interest in the normalization ridge regression method it is necessary to caution the reader to exercise a healthy scepticism, and at least until more evidence has accumulated. For example, though we have shown with reference to four different examples that normalization ridge regression holds its own as a biased, variance-reduced procedure, it is highly likely that how well the method performs compared to OLS and other ridge regression methods will depend on the condition of the data set. But the conditions of the data set which may influence normalization regression results were not identified. Neither could they be identified in this instance. But they probably include the degree of multicollinearity, the signal to noise ratio, and the accuracy of model specification. And there are almost certainly other unknown conditions governing the effectiveness of any ridge regression method. Some answers will be forthcoming from the results of future Monte Carlo simulation experiments. Though the goal is to identify the single, best, all-purpose method for solving the multicollinearity problem, it is likely that such a goal will prove elusive. Perhaps the best we can hope for in the near future is to identify the conditions which justify the use of one biased, reduced-variance approach in preference to others.

Does this mean that we are no closer to the calculation of the optimum k (k_o) than when we started? Probably not. The normalization k we have concluded seems to perform as satisfactorily as five other ridge regression methods. But, as pointed out in section 4.3, the other methods were based on a different criterion. This criterion -- minimizing the mean square error -- despite its six important advantages over OLS regression (section 3.4) seems flawed in three respects. First, the procedure is stochastic when it should be nonstochastic (section 3.4). Second, with the MSE criterion, even when the data set is orthogonal there exists a positive optimal k, though theoretically under this condition k should equal zero. Third, the ridge procedure based on the MSE criterion is multifunctional. It suppresses the variance-inflation due not only to multicollinearity, but also to measurement error, model misspecification and small sample size (sections 4.1 and 4.2). Use of the normalization criterion avoids all three of these flaws. In this respect, then, the normalization ridge method may be advantageous. But to what extent we do not know.

Can we say anything with certainty about the different ridge regression methods? Unfortunately not. We are fairly confident, though not certain, that OLS results are usually erroneous and often uninterpretable when the multicollinearity is high. But we do not know the degree of multicollinearity necessary for the OLS procedure to be erroneous or of dubious value. Of the ridge regression methods themselves we are fairly confident, but not certain, that the Kasarda and Shih and Vinod results are consistently weaker than those of five other methods. Of these five methods the available evidence seems to support the superiority of two of them - the normative and the Hoerl and Kennard iterative. Nevertheless, we pointed out that we had little confidence in the iterative result for the first example, and we expressed reservations about the iterative result in the third example because the estimates were oversuppressed, and therefore, the multicollinearity problem had been overcorrected. We recommended that the tolerance level of the iterative procedure be adjusted accordingly as a way of coping with the first reservation, but not with the second. We were also able to show in terms of the performance indices that of all the methods the normalization one operated in the most consistent manner.

The evidence supported the conclusion that for examples two and four the results of the normative method and iterative method were essentially identical. Thus, for the air pollution problem the four significant factors were mean annual precipitation, the mean January temperature, the percent of the population that is non-white, and air pollution in the form of sulfur dioxide. One interpretation might be that the life expectancies of non-whites living in cool winter climates where year round precipitation levels are high and where sulfur dioxide air pollution is constantly present are lower than for whites. The important general finding is that when controlling for a host of possible factors affecting mortality rates the air pollution variable emerges as a significant factor in its own right. The intuitions of environmentalists and others concerned with promoting healthy, pollution-free environments were confirmed by the analysis; there is little doubt that the presence of air pollution can have negative influences on human life expectancies.

In the employee satisfaction study two factors or supervisory principles emerged as the major determinants of employee satisfaction: how supervisors handled employee complaints and whether supervisors applied company policy fairly. Three other considerations were found to have negligible significance.

As to what type of car to buy in order to obtain the best gas mileage - it all depends on which method of analysis is preferred. On the performance indices the iterative method had the highest rating and according to this method the most important factor is the weight of the vehicle (the lighter the better) followed by the number of carburetor barrels (the fewer the better). Three other factors were also significant and about equal in importance; namely, the number of cylinders (four is better than six or eight), manual transmission, and the number of horsepower. The normalization ridge regression performance was also good. According to this method only two factors make much difference; the vehicle weight and the number of carburetors.

CHAPTER SIX

6.0 CONCLUSION

1. The six harmful effects of multicollinearity on OLS regression estimation were described in Chapter One. In general, the Gauss-Markov property of minimum variance retained by OLS regression is inflated in the presence of multicollinearity and may even become unacceptably large. Further, the OLS model is theoretically unbiased only when the underlying auxiliary model is fully specified. Thus, though unbiasedness and minimum variance are highly desirable properties of any estimating technique, and though theoretically OLS regression retains both properties, it is useful to note with Goldberger (1964, p. 127) that "unbiasedness is not sacred; it may well be reasonable to prefer an estimator that has a small bias and a small variance to one that has no bias but a large variance."

2. Ridge regression analysts have found in fact that the "small" variance property of OLS regression does not necessarily offset the bias when multicollinearity is problematic. Hence, ridge regression was developed as a biased, reduced-variance method to offset this problem. In Chapter Two the sixteen properties of ridge regression were described. In Chapter Three some three distinctive approaches to ridge regression analysis were discussed in detail. Most of the properties of ridge regression are advantageous, or desirable ones. Collectively, they go much of distance toward overcoming the six harmful effects of multicollinearity described in Chapter One.

From property four (2.2.4) one notes that although OLS estimates may have large sampling variance, the ridge technique will attenuate the variance, thereby generating smaller sampling variance. But it is because the OLS estimates may have large sampling variance that they are sensitive to sampling fluctuation. It follows, then, that because ridge regression produces smaller sampling variance that the ridge estimate is less sensitive to sampling fluctuation.

Because ridge regression reduces the variance of the estimates (2.2.4) it is less sensitive to model misspecification than OLS solutions. This is because stochastic error is a source of the total variance of the estimated coefficients (1.4.1). In the fourth place it is noted that the greater the variance of the estimates the less their independence; that is, the higher their correlation the greater the potential for error (1.2.5). Thus, any procedure such as ridge regression which reduces the variance of the estimates will generate more accurate estimates if the biasing parameter is not too large.

Fifth, OLS coefficient vectors are too long when the data set is ill-conditioned; hence, runs the risk of being "off" both in magnitude and direction (1.2.2). But it was shown above (2.2.1) that the ridge procedure necessarily gives a shorter coefficient vector than OLS. Sixth, in the presence of multicollinearity OLS models will not provide accurate prediction, even when the R square is high. However, in section 2.2.11 it was shown that even though ridge regression produces a smaller R square than OLS regression it always provides a more accurate prediction equation if the bias is not too large. The proof of this assertion was given in section 2.2.13.

3. Not all the properties of ridge regression were found to be

advantageous. In fact, the disadvantages could on occasion outweigh the advantages. Most of the disadvantages were discussed in section 3.4 -- especially properties eight (2.2.8), fifteen (2.2.15) and sixteen (2.2.16), which are related to the use of the minimum mean square error criterion. Thus, the ridge regression procedure seemed flawed in four important respects. Fist, an optimum biasing parameter, k_o, cannot be calculated because it depends on unknown population parameters. Thus, it has to be estimated, and numerous different estimating techniques are possible. All give different results.

Second, the biasing parameter, k, is stochastic when it should be nonstochastic. This is because k depends on β (the true coefficient vector) and σ^2 (the variance of the residual) which are, in turn, dependent on the y-variable. Thus, k, depends on the interdependencies between the dependent and independent variables even though the multicollinearity problem arises solely from the interdependencies among the explanatory variables, X's. It is because estimates of the biasing parameter, k, have to be based on X-y dependencies that ridge regression estimation is a stochastic procedure even though in theory it should be nonstochastic. Third, for any regression problem there exists a positive "optimal" k which is true even for an orthogonal system for which the multicollinearity is zero. Thus, ridge regression fails to meet the boundary condition (k=0) for an orthogonal data set. Finally, ridge regression is multifunctional. It not only reduces the variance inflation due to multicollinearity but also reduces variance inflation due to measurement error, model misspecification and small sample size.

4. To overcome these limitations of ridge regression an alternative criterion was proposed called the variance normalization criterion.

The variance mornalization criterion was designed to perform a single task; namely, to reduce variance inflation due to multicollinearity and only nulticollinearity. To achieve this, one has to normalize the variance inflation factor so that is is equal to one. The criterion not only retains the advantages of ridge regression using the minimum mean square error criterion, but also overcomes most of disadvantages of simple ridge regression. Thus, it was shown to give a nonstochastic k; it meets the boundary condition such that for an orthogonal model k=0; and it performs the single task of reducing variance inflation due only to multicollinearity. Further, it was found that the average variance inflation factor could never be suppressed, and that the regression procedure based on the normalization criterion performed more consistently in practice than any of the ridge regression methods based on the minimum mean square error criterion.

5. The normalization ridge regression method, then, seems on the basis of performance comparisons (Chapter Five) to be viable enough. In practice its performance characteristics seemed on a par with those of any other ridge regression method and for this reason might reasonably be considered as an acceptable alternative to standard biased, reduced-variance methods. Nevertheless, there are several factors governing the condition of any data set besides the degree of multicollinearity; for example, the model specification, sample size and signal to noise ratio. It is too early in the normalization ridge regression tradition to make extravegant claims without knowing how these conditions affect normalization ridge regression estimation. Some Monte Carlo simulation experiments would throw light on these matters. Obviously, the normalization ridge regression research is in its infancy; though on the basis of current research it would seem to be a vigorous and lively infant.

REFERENCES

Allen, D.M. The relationship between variable selection and data augmentation and method for prediction. Technometrics, 1974, 16, 125-127.

Blalock, H.M.Jr. Measurement and conceptualization problems: The major obstacle to integrating theory and research. American Sociological Review, 44, 1979, 881-894.

Bulcock, J.W., Lee, W.F., & Luk, W.S. A model of school learning estimated by the variance normalization method of ridge regression analysis. Symposium on Models for the Analysis of Data from the Second IEA Mathematics Study. East Lansing: Michigan State University, January 1979.

Bulcock, J.W., Lee, W.F., & Luk, W.S. The variance normalization method of ridge regression analysis. Princeton, N.J.: ERIC Clearinghouse on Tests, Measurements and Evaluation, Educational Testing Service, 1980 (ERIC ED 178 581).

Bulcock, J.W., Lee, W.F., & Luk, W.S. Normalization regression estimation with application to a nonorthogonal, nonrecursive model of school learning. Canadian Journal of Education (forthcoming 1981).

de Lone, R.H. Small Futures: Children, Inequality, and the Limits of Liberal Reform. New York: Harcourt Brace Jorvanovich, 1979.

Farrar, D.E., & Glauber, R.R. Multicollinearity in regression analysis: The problem revisited. Review of Economics and Statistics, 1967, 92-107.

Fisher, R.A. Statistical Methods for Research Workers. New York: Hafner Publishing Co. 1950 (first published, 1925).

Fisher, R.A. The Design of Experiments. London: Oliver and Boyd, 1935.

Frisch, R. Statistical Confluence Analysis by Means of Complete Regression Systems. Oslo: University Economics Institute, 1934.

Goldberger, A.S. Econometric Theory. New York: Wiley, 1964.

Gorman, J.W., & Toman, R.J. Selection of variables for fitting equations to data. *Technometrics*, 8, 1966, 27-51.

Guilkey, D.K., & Murphy, J.L. Directed ridge regression techniques in case of multicollinearity. *Journal of the American Statistical Association*, 70, 1975, 769-775.

Hemmerle, W.J. An explicit solution for generalized ridge regression. *Technometrics*, 17, 1975, 309-314.

Hemmerle, W.J., & Brantle, T.F. Explicit and constrained generalized ridge estimation. *Technometrics*, 20, 1978, 109-120.

Hickey, R.J., Boyce, D.E., Harner, E.B., & Clelland, R.C. Ecological statistical studies concerning environmental pollution and chronic disease. IEEE Transactions on Geoscience Electronics GE-8, 186-202.

Hocking, R.R. The analysis and selection of variables in linear regression. *Biometrics*, 32, 1976, 1-49.

Hocking, R.R., Speed, F.M., & Lynn, M.J. A class of biased estimators in linear regression. *Technometrics*, 12, 1976, 55-67.

Hoerl, A.E. Optimum solution of many variables equations. *Chemical Engineering Progress*, 55, 1959, 69-78.

Hoerl, A.E. Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58, 1962, 54-59.

Hoerl, A.E. Ridge analysis. *Chemical Engineering Progress Symposium*, 60, 1964, 67-77.

Hoerl, A.E., & Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 1970a, 55-67.

Hoerl, A.E., & Kennard, R.W. Ridge regression: Application to nonorthogonal problems. *Technometrics*, 12, 1970b, 69-82.

Hoerl, A.E., & Kennard, R.W. Ridge regression: Iterative estimation of the biasing parameter. *Communications in Statistics -- Theoretical Methods*, A5(1), 1976, 77-88.

Hoerl, A.E., Kennard, R.W., & Baldwin, K.F. Ridge regression: Some simulation. Communications in Statistics, 4(2), 1975, 105-123.

Intriligator, M.D. Econometric Models, Techniques, and Applications. Englewood Cliffs, N.J.: Prentice-Hall, 1978.

Johnston, J. Econometric Methods. (2nd Edition). New York: McGraw-Hill, 1972.

Kasarda, J.D., & Shih, W.F. Optimal bias in ridge regression approaches to multicollinearity. Sociological Methods and Research, 5, 1977, 461-470.

Keniston, K. All Our Children: The American Family Under Pressure. New York: Harcourt Brace Javanovich, 1977.

King, D. Feedback and reciprocal causation: A Monte Carlo study of the relative small sample properties of small sample estimators. Unpublished doctoral dissertation, University of North Carolina, 1974.

Kish, L. Representation, randomization, and control. In H.M. Blalock, Jr. et.al. Mathematical Models and International Perspectives on Mathematical Sociology. New York: Academic Press, 1975.

Land, K.C., & Felson, M. Sensitivity analysis of arbitrarity identified simultaneous-equation models. Sociological Methods and Research, 6, 1978, 238-307.

Lave, L.B., & Seskin, E.P. Air pollution and human health. Science, 169, 1970, 723-733.

Lawless, J.F., & Wang, P. A simulation study of ridge and other regression estimators. Communications in Statistics -- Theoretical Methods, A5(4), 1976, 306-323.

Lee, W.F. Model estimation using ridge regression with the variance normalization criterion. (Interim Report No.2, Grant No. S76-0645, The Social Sciences and Humanities Research Council of Canada). St. John's, Newfoundland: Memorial University of Newfoundland, 1979.

Lindley, D.V., & Smith, A.F.M. Bayes estimates for the linear model. Journal of the Royal Statistical Society, Series B, 34, 1972, 1-41.

Maddala, G.S. Econometrics. New York: McGraw-Hill, 1977.

Mason, R.L., Gunst, R.F., & Webster, J.T. Regression analysis and problems of multicollinearity. Communications in Statistics, 4(3), 1975, 277-292.

McCabe, G.P.Jr. Evaluation of regression coefficient estimates using α-acceptability. Technometrics, 20, 1978, 131-139.

McDonald, G.C., & Schwing, R.C. Instabilities of regression estimates relating air pollution to mortality. Technometrics, 15, 1973, 463-481.

McDonald, G.C., & Galarneau, D.I. A Monte Carlo evaluation of some ridge-type estimators. Journal of the American Statistical Association, 70, 1975, 407-416.

Marquardt, D.W. Generalized inverse, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics, 12, 1970, 591-612.

Marquardt, D.W., & Snee, R.D. Ridge regression in practice. The American Statistician, 29, 1975, 3-20.

Mill, J.S. A System of Logic, Raciocinative and Inductive: Book III. New York: Harper and Brothers, 1891.

Obenchain, R.L. Ridge analysis following a preliminary test of the shrunken hypothesis. Technometrics, 17, 1975, 431-445.

Ogbu, J.U. Minority Education and Caste: The American System in Cross-Cultural Perspective. New York: Academic Press, 1978.

Pindyck, R.S., & Rubinfeld, D.L. Econometric Methods and Economic Forecasts. New York: McGraw-Hill, 1976.

Price, B. Ridge regression: Application to non-experimental data. Psychological Bulletin, 84, 1977, 759-766.

Stein, C. Inadmissability of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium on Mathematical Statistical Problems, 1, 1955, 197-206.

Vinod, H.D. Application of new ridge regression methods to a study of Bell system scale economics. Journal of the American Statistical Association, 71, 1976, 835-841.

Wannacott, R.J., & Wannacott, J.H. Econometrics. New York: Wiley, 1969.

White, V.J. A general computer program for econometric methods: SHAZAM. Econometrica, 46, 1978, 239-240.

Widder, D.V. Advanced Calculus. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

APPENDIX A. PROGRAM SNR

A.1 DESCRIPTION

The program SNR (Simple Normalization Regression) is listed below. Though for most users it will be self explantary, notice is given that it is only semi-modulated and in its present format is machine specific. The University of Alberta operates MTS (Michigan Terminal System) on an Amdahl 470V/7 mainframe.

The SNR program performs the following analyses:

(i) Ordinary Least Squares Regression
(ii) Normalization Ridge Regression
(iii) Kasarda and Shih's Ridge Regression
(iv) Hocking, Speed and Lynn's Ridge Regression
(v) Hoerl, Kennard and Baldwin's Ridge Regression
(vi) Lawless and Wang's Ridge Regression
(vii) Hoerl and Kennard's Iterative Ridge Regression
(viii) Vinod's Ridge Regression
(ix) Hoerl and Kennard's Ridge Trace Method of Ridge Regression

Readers wishing more information about this program may contact Dr. W.S. Luk, c/o Computing Science, Rm. 7316, Simon Fraser University, Burnaby, B.C., Canada, V5A 1S6.

```
C
C     SIMPLE REGRESSION (KN, KS, KK, KH, KW & KV)
C
C    GIVEN: X'X (XPIX) - A P X P MATRIX
C           X'Y (XPIYY) - A P-VECTOR
C           P - SCALAR, INTEGER, MAXIMUM 10
C
C    DATA INPUT FORMAT:
C    1ST CARD:
C      COL.1 - 2: VALUE OF P, RIGHT JUSTIFIED
C      COL.3 - 7: VALUE OF N, RIGHT JUSTIFIED
C    2ND - (P+1)TH CARD: (FOR THE XPIX MATRIX)
C      EACH CARD CONTAINS A ROW OF P REAL NUMBERS. EVERY 7 COLUMNS OF A
C      CARD STARTING FORM THE 1ST UP TO THE (7P)TH COLUMN IS OCCUPIED
C      BY A REAL NUMBER IN THE FORM S9.9999 WHERE THE SIGN S CAN BE
C      IF IT IS POSITIVE.
C    (P+2)TH CARD: (FOR THE XPIYY MATRIX)
C      THIS CARD AGAIN CONTAINS A ROW OF P REAL NUMBERS IN THE SAME
C      FORMAT AS ABOVE.
C
C    NOTE:
C       THIS PROGRAM REQUIRES FOUR IMSL SUBROUTINES:
C     1. VCVTFS: TO CONVERT A MATRIX INTO A FORMAT ACCEPTABLE TO EIGRS
C     2. EIGRS: TO FIND EIGENVALUES AND EIGENVECTORS OF A MATRIX
C     3. LINV2F: TO FIND THE INVERSE OF A MATRIX
C     4. MDFD: TO FIND THE SIGNIFICANCE LEVEL OF A F-DISTRIBUTION
C           GIVEN THE VALUE OF F
C
      IMPLICIT REAL*8(A-H),INTEGER*4(I-J),REAL*8(K-Z)
      INTEGER P,IJOB/2/,N,NTIMES,NB
      INTEGER BLANK/'    '/,AKS/'(KS)'/,AKN/'(KN)'/,AKV/'(KV)'/
      INTEGER AKK/'(KK)'/,AKH/'(KH)'/,AKW/'(KW)'/
      REAL*8 XPIX(20,20),A(210),LAMBDA(20),PMATRX(20,20),ALPHA(20),
     +       ALPHA$(20),BETA(20),BETA$(20),XPIY(20),XXINV(20,20),
     +       VIF(20,20),WK(475),KV(20),XPIXK(20,20),XXKINV(20,20),
     +      XPI$Y(20),BETAP(20),BETA$P(20),
     +       PERTXX(20,20),PXXINV(20,20),PXXKIN(20,20),PXXK(20,20),
     +       RHO(20),TT(20),R(20,20),RHO$(20),PLAMDA(20),R$(20,20),
     +       RTCHAN(20),T$(20)
 10   FORMAT(I2,I5)
 11   FORMAT(10F7.4)
 12   FORMAT('-INPUT VALUES: '/' ----- ------')
 13   FORMAT('0X''X MATRIX: ')
 14   FORMAT('   ',10F8.4/'   ',10F8.4)
 15   FORMAT('0X''Y VECTOR: ',10F8.4/'             ',10F8.4)
```

```
16    FORMAT('OBETA :',10E12.4/'          ',10E12.4)
17    FORMAT('ORZERO SQUARE: ',E12.4)
18    FORMAT('OBETA*: ',10E12.4/'          ',10E12.4)
19    FORMAT('OALPHA*: ',10E12.4/'           ',10E12.4)
20    FORMAT('OPHI: ',E12.4)
21    FORMAT('OMPHI: ',E12.4)
22    FORMAT('OD: ',E12.4)
23    FORMAT('OSIGMA SQUARE: ',E12.4)
24    FORMAT('OR SQUARE: ',E12.4)
25    FORMAT('1',5X,'K',13X,'AVVIF',8X,'EQ.D',8X,'RMSE',8X,'RVAR',
     *        8X,'RBIAS',8X,'EFT',8X,'SQBIAS',8X,'RSD')
26    FORMAT(' ',E12.4,A4,9E12.4)
27    FORMAT('OK: ',E12.4)
28    FORMAT(' ****************************************************')
29    FORMAT(' *                                                  *')
30    FORMAT(' *  SIMPLE REGRESSION (KN, KS, KK, KH, KW & KV)     *')
31    FORMAT('OP: ',I2,', N: ',I4)
32    FORMAT('-OUTPUT VALUES:'/' ------ ------')
33    FORMAT('ODMAX =',E12.4)
34    FORMAT('1HOCKING''S METHOD:'/' --------- ------')
35    FORMAT('1OLS:'/' ---')
36    FORMAT('1NORMALIZATION METHOD:'/' ------------- ------')
37    FORMAT('1VINOD''S METHOD:',I2,' VALUES'/' ------- ------')
38    FORMAT('1SIMPLE REGRESSION (KN,KV&KS): X(',I2,') REGRESSES ON',
     #      ' X(1) TO X(',I2,')')
39    FORMAT(10E12.4)
40    FORMAT('1HKB''S METHOD:'/' ----- ------')
41    FORMAT('1LAWLESS & WANG''S METHOD:'/
     #      ' ------- - ------ ------')
42    FORMAT('OS.RATIO(2):',E12.4)
43    FORMAT('OS.RATIO(3):',E12.4,'/',E11.4,' = ',E12.4)
44    FORMAT('ODC:',E12.4)
45    FORMAT('1ITERATIVE METHOD:'/' --------- ------')
70    FORMAT('OLAMBDA =',10E12.4/'           ',10E12.4)
71    FORMAT('OP MATRIX:')
72    FORMAT('OVMAX :',E12.4)
73    FORMAT('ODMAX:',E12.4)
74    FORMAT('OX''X INVERSE:')
75    FORMAT('OSQB*:',E12.4)
76    FORMAT('ORBIAS*:',E12.4)
77    FORMAT('OEFT*',E12.4)
78    FORMAT('ORMSE*:',E12.4)
79    FORMAT('OVIF MATRIX:')
80    FORMAT('OVIF MAX:',E12.4)
81    FORMAT('OEQDMAX:',E12.4)
```

```
  82  FORMAT('0ALPHA:',10E12.4/'       ',10E12.4)
  83  FORMAT('0KS:',E12.4)
  84  FORMAT('0PERTURBED X''X:')
  85  FORMAT('0PRTBED BETA:',10E12.4/'            ',10E12.4)
  86  FORMAT('0S.RATIO(1):',E12.4)
  87  FORMAT('0EFFIC:',E12.4)
  88  FORMAT('0PRTBD BETA*',10E12.4/'            ',10E12.4)
  89  FORMAT('1KASARDA''S METHOD:'/' --------- ------')
  90  FORMAT('0RTCHAN: ',10E12.4/'        ',10E12.4)
 501  FORMAT('1RIDGE TRACE METHOD:'/' ----- ----- ------')
 502  FORMAT('- K     BETA*')
 503  FORMAT(' ',F4.2,'   ',10E12.4/'        ',10E12.4)
 504  FORMAT('0EIGENVALUES OF PERTURBED X''X:')
 505  FORMAT('0RHO:',10E12.4/'     ',10E12.4)
 506  FORMAT('0T:',10E12.4/'  ',10E12.4)
 507  FORMAT('0R MATRIX:')
 508  FORMAT('0S.D* :',10E12.4/'       ',10E12.4)
 509  FORMAT('0R* MATRIX:')
 510  FORMAT('0F:',E12.4,',  ACCEPTANCE LEVEL',F6.3)
 511  FORMAT('0T*:',10E12.4/'    ',10E12.4)
  512  FORMAT('0AVVIF:',E12.4)
  513  FORMAT('0EQD:',E12.4)
  514  FORMAT('0RVAR:',E12.4)
C
C     INPUT DATA
C
      DO 995 IRUN=1,10
      READ(5,10,END=999) P,N
      DO 240 I=1,P
 240  READ(5,11) (XPIX(I,J),J=1,P)
      READ(5,11) (XPIY(I),I=1,P)
C
C     FIND PERTXX (PERTURBED X''X)
C
      DELTAR=0.01
      MAX=0.0
      DO 123 I=1,P
      DO 123 J=1,P
      PERTXX(I,J)=XPIX(I,J)
      IF(MAX.GE.DABS(XPIX(I,J)).OR.I.EQ.J) GO TO 123
      MAX=DABS(XPIX(I,J))
      I1=I
      J1=J
  123 CONTINUE
      IF (PERTXX(I1,J1).GE.0.)  PERTXX(I1,J1)=PERTXX(I1,J1)+DELTAR
      IF (PERTXX(I1,J1).LT.O.)  PERTXX(I1,J1)=PERTXX(I1,J1)-DELTAR
      PERTXX(J1,I1)=PERTXX(I1,J1)
C
C     PRINT OUT THE INPUT DATA
C
      WRITE(6,12)
      WRITE(6,31) P,N
      WRITE(6,13)
      DO 241 I=1,P
  241  WRITE(6,14) (XPIX(I,J),J=1,P)
      WRITE(6,15) (XPIY(I),I=1,P)
      WRITE(6,84)
      DO 124 I=1,P
 124  WRITE(6,14) (PERTXX(I,J),J=1,P)
      WRITE(6,32)
      IP=20
      CALL VCVTFS(PERTXX,P,IP,A)
      CALL EIGRS(A,P,IJOB,PLAMDA,PMATRX,IP,WK,IER)
      WRITE(6,504)
      WRITE(6,39) (PLAMDA(I),I=1,P)
      CALL VCVTFS(XPIX,P,IP,A)
      CALL EIGRS(A,P,IJOB,LAMBDA,PMATRX,IP,WK,IER)
C
C      ORINDARY LEAST SQUARE METHOD (OLS)
C
      IDGT=O
      CALL LINV2F(XPIX,P,IP,XXINV,IDGT,WK,IER)
      VMAX=-99999.O
      DO 97 I=1,P
  97  IF (XXINV(I,I).GT.VMAX) VMAX=XXINV(I,I)
      WRITE(6,70) (LAMBDA(I),I=1,P)
      WRITE(6,71)
      DO 98 I=1,P
  98  WRITE(6,39) (PMATRX(I,J),J=1,P)
      WRITE(6,74)
      DO 262 I=1,P
 262  WRITE(6,39) (XXINV(I,J),J=1,P)
      WRITE(6,72) VMAX
      IDGT=O
      CALL LINV2F(PERTXX,P,IP,PXXINV,IDGT,WK,IER)
      WRITE(6,35)
C
C     CALCUALTE BETA
C
      DO 243 I=1,P
      BETA(I)=0.0
      DO 243 J=1,P
  243  BETA(I)=BETA(I)+XXINV(I,J)*XPIY(J)
      WRITE(6,16) (BETA(I),I=1,P)
C
C     CALCULATE RATE OF CHANGE
C
      DO 225 I=1,P
      RTCHAN(I) = 0.0
      DO 225 J=1,P
 225  RTCHAN(I)=RTCHAN(I)+(-XXINV(I,J))*BETA(J)
      WRITE(6,90)(RTCHAN(I),I=1,P)
C
C     CALCULATE PERTURBED BETA
C
      DO 125 I=1,P
      BETAP(I)=0.0
      DO 125 J=1,P
  125 BETAP(I)=BETAP(I)+PXXINV(I,J)*XPIY(J)
      WRITE(6,85) (BETAP(I),I=1,P)
C
C     CALCULATE ALPHA
C
      DO 108 I=1,P
      XPI$Y(I)=0.0
      DO 109 J=1,P
 109  XPI$Y(I)=XPI$Y(I)+PMATRX(J,I)*XPIY(J)
 108  ALPHA(I)=XPI$Y(I)/LAMBDA(I)
      WRITE(6,82) (ALPHA(I),I=1,P)
C
C     CALCULATE RZEROSQUARE
C
      RZERO2=0.0
      DO 247 I=1,P
 247  RZERO2=RZERO2+BETA(I)*XPIY(I)
      WRITE(6,17) RZERO2
C
C     CALCULATE TR(X'XINV)
C
      TRXXI=0.0
      DO 249 I=1,P
 249  TRXXI=TRXXI+(1./LAMBDA(I))
C
C     CALCULATE D
C
```

```
      D=(2./3.1416)*(DATAN(TRXXI/DFLOAT(P)-1.))
      WRITE(6,22) D
C
C     CALCULATE DMAX
C
      DMAX=(2./3.1416)*DATAN(VMAX-1.)
      WRITE(6,33) DMAX
C
C     CALCULATE D.C
C
      DC=(2./3.1416)*DATAN(1./LAMBDA(1)-1.)
      WRITE(6,44) DC
C
C     CALCULATE SIGMA SQUARE
C
      SIGMA2=(1.-RZERO2)/DFLOAT(N-P)
      WRITE(6,23) SIGMA2
C
C     CALCULATE RHO, T
C
      DO 133 I=1,P
      RHO(I)=DSQRT(SIGMA2*XXINV(I,I))
  133 TT(I)=BETA(I)/RHO(I)
      WRITE(6,505) (RHO(I),I=1,P)
      WRITE(6,506) (TT(I),I=1,P)
C
C     CALCULATE R MATRIX
      WRITE(6,507)
      DO 134 I=1,P
      DO 135 J=1,P
  135 R(I,J)=XXINV(I,J)/(DSQRT(XXINV(I,I))*DSQRT(XXINV(J,J)))
      WRITE(6,39) (R(I,J),J=1,P)
  134 CONTINUE
C
C     FIND ALL KV'S
C
      CALL OPTIMK(LAMBDA,P,KV,IHEAD)
      IH=IHEAD+6
      IHH=0
C
C     FOR EACH OF KN, KS, KK, KH, KW & KV, CALCULATE RIDGE STATISTICS
C
      DO 59 IJ=1,IH
      IF (IJ.GT.1) GO TO 141
      CALL NORMLK(LAMBDA,P,KN)
      K=KN
      WRITE(6,36)
      GO TO 57
  141 IF (IJ.GT.2) GO TO 142
      CALL KASRDS(LAMBDA,ALPHA,SIGMA2,P,KS)
      K=KS
      WRITE(6,89)
      GO TO 57
  142 IF (IJ.GT.3) GO TO 143
      CALL HOCKIN(LAMBDA,ALPHA,SIGMA2,P,KK)
      WRITE(6,34)
      GO TO 57
  143 IF (IJ.GT.4) GO TO 144
      CALL HKB(ALPHA,SIGMA2,P,KH)
      K=KH
      WRITE(6,40)
      GO TO 57
  144 IF (IJ.GT.5) GO TO 145
      CALL LAWANG(LAMBDA,ALPHA,SIGMA2,P,KW)
      K=KW
      WRITE(6,41)
      GO TO 57
  145 IF (IJ.GT.6) GO TO 146
      CALL ITER(ALPHA,LAMBDA,SIGMA2,TRXXI,P,KI)
      K=KI
      WRITE(6,45)
      GO TO 57
  146 IF (IJ.EQ.7) WRITE(6,37) IHEAD
      IHH=IHH+1
      K=KV(IHH)
C
C     CALCULATE BETA*
C
   57 WRITE(6,27) K
      DO 244 I=1,P
      DO 244 J=1,P
      XPIXK(I,J)=XPIX(I,J)
      IF (I.EQ.J) XPIXK(I,J)=XPIX(I,J)+K
  244 CONTINUE
      IDGT=0
      IP=20
      CALL LINV2F(XPIXK,P,IP,XXKINV,IDGT,WK,IER)
      DO 245 I=1,P
      BETA$(I)=0.0
      DO 245 J=1,P
  245 BETA$(I)=BETA$(I)+XXKINV(I,J)*XPIY(J)                             00002970
      WRITE(6,18) (BETA$(I),I=1,P)                                      00002980
C                                                                       00002990
C     CALCULATE PERTURBED BETA$                                         00003000
C                                                                       00003010
      DO 126 I=1,P                                                      00003020
      DO 126 J=1,P                                                      00003030
      PXXK(I,J)=PERTXX(I,J)                                             00003040
      IF (I.EQ.J) PXXK(I,J)=PERTXX(I,J)+K                               00003050
  126 CONTINUE                                                          00003060
      IDGT=0                                                            00003070
      CALL LINV2F(PXXK,P,IP,PXXKIN,IDGT,WK,IER)                         00003080
      DO 127 I=1,P                                                      00003090
      BETA$P(I)=0.0                                                     00003100
      DO 127 J=1,P                                                      00003110
  127 BETA$P(I)=BETA$P(I)+PXXKIN(I,J)*XPIY(J)                           00003120
C                                                                       00003140
C     CALCULATE ALPHA*                                                  00003150
C                                                                       00003160
      DO 246 I=1,P                                                      00003170
      ALPHA$(I)=0.0                                                     00003180
      DO 246 J=1,P                                                      00003190
  246 ALPHA$(I)=ALPHA$(I)+PMATRX(J,I)*BETA$(J)                          00003200
C                                                                       00003220
C     CALCULATE PHI                                                     00003230
C                                                                       00003240
      PHI=0.0                                                           00003250
      DO 248 I=1,P                                                      00003260
  248 PHI=PHI+ALPHA$(I)**2/LAMBDA(I)                                    00003270
      PHI=PHI*K*K                                                       00003280
      WRITE(6,20) PHI                                                   00003290
C                                                                       00003300
C     CALCULATE MPHI                                                    00003310
C                                                                       00003320
      MPHI=PHI/(1.-RZERO2)                                              00003330
      WRITE(6,21) MPHI                                                  00003340
C                                                                       00003350
C     CALCULATE R SQUARE                                                00003360
C                                                                       00003370
      R2=0.0                                                            00003380
      DO 50 I=1,P                                                       00003390
      T=0.0                                                             00003400
      DO 61 J=1,P                                                       00003410
   61 T=T+XPIX(I,J)*BETA$(J)                                            00003420
      R2=R2+BETA$(I)*T                                                  00003430
   50 CONTINUE
      WRITE(6,24) R2
C
C     CALCULATE VIF AND VIF MAX
C
      DO 65 I=1,P
      DO 65 J=1,P
   65 VIF(I,J)=0.0
      DO 66 I=1,P
      DO 66 J=1,P
      DO 66 IK=1,P
   66 VIF(I,J)=VIF(I,J)+PMATRX(I,IK)*(LAMBDA(IK)/(LAMBDA(IK)+K)**
     #          2)*PMATRX(J,IK)
      WRITE(6,79)
      DO 67 I=1,P
   67 WRITE(6,39) (VIF(I,J),J=1,P)
      VIFMAX=-99999.0
      DO 68 I=1,P
   68 IF (VIF(I,I).GT.VIFMAX) VIFMAX=VIF(I,I)
C
C     CALCULATE RHO*
C
      DO 136 I=1,P
      RHO$(I)=DSQRT(SIGMA2*VIF(I,I))
  136 T$(I)=BETA$(I)/RHO$(I)
      WRITE(6,508) (RHO$(I),I=1,P)
      WRITE(6,511) (T$(I),I=1,P)
C
C     CALCULATE R* MATRIX
C
      WRITE(6,509)
      DO 137 I=1,P
      DO 138 J=1,P
  138 R$(I,J)=VIF(I,J)/(DSQRT(VIF(I,I))*DSQRT(VIF(J,J)))
  137 WRITE(6,39) (R$(I,J),J=1,P)
C
C     CALCULATE F & ITS SIGNIFICANCE LEVEL
C
      F=0.0
      DO 139 I=1,P
      DO 139 J=1,P
  139 F=F+(BETA$(I)-BETA(I))*XPIX(I,J)*(BETA$(J)-BETA(J))
      F=F/(DFLOAT(P)*SIGMA2)
      I1=P
      I2=N-P
      CALL MDFD(F,I1,I2,SIGLVL,IER)                                     00003880
      SIGLVL=1.-SIGLVL                                                  00003890
      WRITE(6,510) F,SIGLVL                                             00003900
C                                                                       00003910
C     CALCULATE EQ.DMAX                                                 00003920
C                                                                       00003930
      EQDMAX=(2./3.1416)*DATAN(VIFMAX-1.)                               00003940
      WRITE(6,81) EQDMAX                                                00003950
C                                                                       00003960
C     CALCULATE S.RATIO.1                                               00003970
C                                                                       00003980
      S1=0.0                                                            00003990
      DO 128 I=1,P                                                      00004000
  128 S1=S1+(BETA$(I)-BETA$P(I))**2                                     00004010
      S2=0.0                                                            00004020
      DO 129 I=1,P                                                      00004030
  129 S2=S2+(BETA(I)-BETAP(I))**2                                       00004040
      SRATIO=S1/S2                                                      00004050
      WRITE(6,86) SRATIO                                                00004060
      TRVIF=0.0                                                         00004770
      DO 104 I=1,P                                                      00004780
  104 TRVIF=TRVIF+(LAMBDA(I)/((LAMBDA(I)+K)**2))                        00004790
      AVVIF=(1./DFLOAT(P))*TRVIF                                        00004800
      T=0.0                                                             00004810
      DO 53 I=1,P                                                       00004820
   53 T=T+BETA(I)**2                                                    00004830
      RVAR=1.-(TRVIF/TRXXI)                                             00004831
      EQD=(2./3.1416)*(DATAN(TRVIF/DFLOAT(P)-1.))                       0  0 8
       WRITE(6,514) RVAR                                                00004972
  59   CONTINUE                                                         00004980
C                                                                       00004990
C      RIDGE TRACE METHOD                                               00005000
C                                                                       00005010
      WRITE(6,501)                                                      00005020
      WRITE(6,502)                                                      00005030
      K=0.0                                                             00005040
      DO 130 NTIMES=1,19                                                00005050
      IF (NTIMES.LE.10) K=K+0.01                                        00005060
      IF (NTIMES.GT.10) K=K+0.1                                         00005070
      DO 131 I=1,P                                                      00005080
      DO 131 J=1,P                                                      00005090
      XPIXK(I,J)=XPIX(I,J)                                              00005100
      IF (I.EQ.J) XPIXK(I,J)=XPIX(I,J)+K                                00005110
  131 CONTINUE                                                          00005120
      IDGT=0                                                            00005130
      CALL LINV2F(XPIXK,P,IP,XXKINV,IDGT,WK,IER)
      DO 132 I=1,P
      BETA$(I)=0.0
      DO 132 J=1,P
  132 BETA$(I)=BETA$(I)+XXKINV(I,J)*XPIY(J)
      WRITE(6,503) (K,(BETA$(I),I=1,P))
  995 CONTINUE
  999 STOP
      END
      SUBROUTINE OPTIMK(LAMBDA,P,KSTACK,IHEAD)
C
C     THIS ROUTINE CALCULATES OPTIMAL VALUE OF K
C     ACCORDING TO VIOND'S METHOD
C
      IMPLICIT REAL*8(A-H),INTEGER*4(I-J),REAL*8(K-Z)
      INTEGER SWITCH,P,N,KINCRE
      REAL*8 LAMBDA(20),KSTACK(20),ISRM
      A=-4.0
      DELTAA=0.1
      MISRM=1000.00
      IHEAD=0
      SWITCH=0
      KP1=0.0
      KP2=0.0
      DO 86 KINCRE=1,61
      K=10**A
      T=0.0
      DO 92 I=1,P
   92 T=T+LAMBDA(I)/(LAMBDA(I)+K)**2
      ISRM=0.0
      DO 90 I=1,P
   90 ISRM=ISRM+((P*LAMBDA(I)/(LAMBDA(I)+K)**2)/T-1.)**2
      IF ((ISRM.GT.MISRM).AND.(SWITCH.EQ.1)) GO TO 94
      IF ((ISRM.GT.MISRM).AND.(SWITCH.EQ.0)) GO TO 97
   97 KP1=KP2
      KP2=K
      GO TO 96
   94 IHEAD=IHEAD+1
      KSTACK(IHEAD)=KP1
      SWITCH=0
   96 MISRM=ISRM
      A=A+DELTAA
   86 CONTINUE
      DO 100 J=1,IHEAD
      K=KSTACK(J)
      MISRM=1000.00
      SWITCH=0
      KP1=0.0
      DO 101 KINCRE=1,1000
      T=0.0
      DO 102 I=1,P
  102 T=T+LAMBDA(I)/(LAMBDA(I)+K)**2
      ISRM=0.0
      DO 103 I=1,P
  103 ISRM=ISRM+((P*LAMBDA(I)/(LAMBDA(I)+K)**
      SWITCH=1
      KP1=K
      GO TO 106
C104  WRITE (6,38) KP1,MISRM
C 38  FORMAT('1',5X,'K :  ',E12.4,5X,'MISRM :  ',E12.4)
  104 KSTACK(J)=KP1
      GO TO 100
  106 MISRM=ISRM
      K=K+0.01
  101 CONTINUE
  100 CONTINUE
   99 RETURN
      END
      SUBROUTINE NORMLK(LAMBDA,P,KN)
C
C     THIS ROUTINE CALCULATE NORMALIZED K SUCH THAT
C
C         ___
C         >__   (LAMBDA(I)/(LAMBDA(I)+K)**2)=P
C         I=1,P
C
      REAL*8 LAMBDA(20),K,KN,MIN,LP,D&        &
      MIN=99999.0
      K=0.0
      DO 42 J=1,999
      K=K+0.001
      LP=0.0
      DO 60 I=1,P
   60 LP=LP+LAMBDA(I)/(LAMBDA(I)+K)**2
      IF (DABS(LP-DFLOAT(P)).GT.MIN) GO TO 42
      MIN=DABS(LP-DFLOAT(P))
      KN=K
   42 CONTINUE
      RETURN
      END
      SUBROUTINE KASRDS(LAMBDA,ALPHA,SIGMA2,P,KS)
```

```
C
C     THIS ROUTINE CALCULATES K SUCH THAT MSE IS MINIMAL
C
      REAL*8 LAMBDA(20),ALPHA(20),K,KS,T1,T2,MSE,T
      INTEGER P
      MSE=99999.0
      K=0.0
      DO 120 I=1,1000
      K=K+0.001
      T1=0.0
      DO 121 J=1,P
  121 T1=T1+LAMBDA(J)/(LAMBDA(J)+K)**2
      T2=0.0
      DO 122 J=1,P
  122 T2=T2+ALPHA(J)**2/(LAMBDA(J)+K)**2
      T=SIGMA2*T1+K*K*T2
      IF (MSE.LE.T) GO TO 120
      MSE=T
      KS=K
  120 CONTINUE
      RETURN
      END
      SUBROUTINE HOCKIN(LAMBDA,ALPHA,SIGMA2,P,KK)
C
C     HOCKING'S METHOD TO CALCULATE K
C
      INTEGER P
      REAL*8 LAMBDA(P),ALPHA(P),SIGMA2,S1,S2,KK
      S1=0.0
      DO 11 I=1,P
   11 S1=S1+(LAMBDA(I)*ALPHA(I))**2
      S2=0.0
      DO 12 I=1,P
   12 S2=S2+LAMBDA(I)**2*ALPHA(I)**4
      KK=SIGMA2*S1/S2
      RETURN
      END
      SUBROUTINE HKB(ALPHA,SIGMA2,P,KH)
C
C     HKB'S METHOD TO CALCULATE K
C
      INTEGER P
      REAL*8 ALPHA(P),SIGMA2,KH
      KH=0.0
      DO 11 I=1,P
   11 KH=KH+(ALPHA(I)**2)
      KH=DFLOAT(P)*SIGMA2/KH
      RETURN
      END
      SUBROUTINE LAWANG(LAMBDA,ALPHA,SIGMA2,P,KW)
C
C     LAWLESS & WANG'S METHOD TO CALCULATE K
C
      INTEGER P
      REAL*8 ALPHA(P),LAMBDA(P),SIGMA2,KW
      KW=0.0
      DO 11 I=1,P
   11 KW=KW+(LAMBDA(I)*ALPHA(I)**2)
      KW=DFLOAT(P)*SIGMA2/KW
      RETURN
      END
      SUBROUTINE ITER(ALPHA,LAMBDA,SIGMA2,TRXXI,P,KI)
C
C    THIS SUBROUTINE CALCULATES KI BY THE ITERATIVE METHOD
C
      IMPLICIT REAL*8(A-H),INTEGER*4(I-J),REAL*8(K-Z)
      INTEGER P
      REAL*8 ALPHA(P),LAMBDA(P),ALPAIT(20),ALPHAS(P),LAMALP(20)
      A=TRXXI/DFLOAT(P)
      T=-1.30*DLOG(A)
      DELTA=DFLOAT(20)*DEXP(T)
      PSIGMA=DFLOAT(P)*SIGMA2
      SUM=0.0
      DO 10 I=1,P
   10 SUM=SUM+ALPHA(I)**2
      KT=PSIGMA/SUM
      DO 40 I=1,P
   40 LAMALP(I)=LAMBDA(I)*ALPHA(I)
      DO 20 J=1,1000
      DO 30 I=1,P
   30 ALPAIT(I)=LAMALP(I)/(LAMBDA(I)+KT)
      SUM=0.0
      DO 50 I=1,P
   50 SUM=SUM+ALPAIT(I)**2
      KTI=PSIGMA/SUM
      TOL=(KTI-KT)/KT
      IF(TOL.LE.DELTA) GO TO 100
      KT=KTI
   20 CONTINUE
```

```
 100  KI=KT                                                                  00006719
      DO 60 I=1,P                                                            00006720
      END                                                                    00006723
  60  ALPHAS(I)=ALPAIT(I)                                                    00006721
      RETURN                                                                 00006722
```